UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Characterizing the co-evolution of protein-protein interactions

Permalink

https://escholarship.org/uc/item/6g79z0xx

Author

Goh, Chern-Sing,

Publication Date

2002

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Characterizing the Co-Evolution of Protein-Protein Interactions

by

Chern-Sing Goh

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

LIBRARY

Copyright 2002
by
Chern-Sing Goh

Dedicated to

My Mother, Siew Eng Goh
My Father, Peng Beng Goh

And to My Dear Jure

USE LIBRARY

Acknowledgements

I thank Fred Cohen, my graduate advisor, for his wisdom and guidance during the past four years. Fred has been an excellent mentor, providing valuable suggestions yet also allowing me the freedom to pursue my own research interests. I would also like to thank the other members of my thesis committee who helped to direct my research. Patricia Babbitt has been a wonderful source of advice and inspiration in both scientific pursuits and career goals; and Henry Bourne has contributed to an invaluable part of my training by encouraging me to always think beyond the boundaries. Thank you.

I am grateful to other UCSF faculty who have been instrumental in the shaping of my career. I thank Tom Ferrin, Kathy Andriole, Robert Fletterick, Peter Walter, Ida Sim, Teri Klein, Conrad Huang, Jorge Oksenberg, and Mark Segal for their thoughtful advice and contributions over the years.

It has been a privilege to be able to know and work with very intelligent and talented people at UCSF. These include my co-authors Dirk Walther, Andrew Bogan, and Marcin Joachimiak. I would also like to thank Jonathan Blake, John-Marc Chandonia, Andrew Wallace, Elaine Meng, Tim Burkoth, Anthony Lau, Cedric Govaerts, Alex Schnoes, and Florence Horn, who have been wonderful friends and contributed their time and knowledge to interesting and thought-provoking discussions. I have also enjoyed time spent with my BMI classmates who include Rey Banatao, Jose Haresco, Liping Zhang, Jing Zhu, Chris Kingsley, Courtney Harper, Ben Polacco, Jay Choi, and Lawrence Lee. Julie Ransom and Ginger Valen deserve special thanks for their excellent administrative support and for making things run smoothly and efficiently. I am grateful to the UCSF Progra

fellowship to me.

I have been

thank my parents, I

my educational dev

pillar of strength ar

to the UCSF Program in Quantitative Biology for bestowing a Burroughs Wellcome fellowship to me.

I have been both lucky and privileged to be surrounded by a wonderful family. I thank my parents, Peng Beng Goh and Siew Eng Goh, for their loving support throughout my educational development and life. Finally, thank you my dear Jure, for being my pillar of strength and encouragement always.

UCSF LIBRARY

Characte

Protein-protein
controlling vari
pathways. Due
genome sequenc
useful in analyzi
this thesis, severa
characterize and
approach is devel
that interact. Thi
that they co-evolv
proteins with unc
demonstrates how
evolution of chemo
Identifying possibl
cytomegalovirus (C
the complex interpl
mechanisms control
In Chapter 3, the
domain interactions
degree of co-evoluti
and/or domains th

Characterizing the Co-Evolution of Protein-Protein Interactions

By Chern-Sing Goh

Protein-protein interactions play crucial roles in many biological systems by controlling various processes involved in metabolic, signaling, and regulatory pathways. Due to the wealth of experimental information collected by recent genome sequencing efforts, computational techniques have become increasingly useful in analyzing large datasets to provide inferences about protein function. In this thesis, several computational approaches are presented that can be utilized to characterize and identify protein-protein interactions. In Chapter 1, a novel approach is developed for quantifying the co-evolution between two protein families that interact. This technique is applied to chemokines and their receptors to show that they co-evolve. Using this analysis, inferences about the binding partners for proteins with uncharacterized binding specificities can be made. Chapter 2 demonstrates how the co-evolutionary analysis can be used to study the putative co-evolution of chemokines and chemokine receptors of both human and viral origin. Identifying possible interactions between the human cellular and the human cytomegalovirus (CMV)-encoded chemokines and chemokine receptors and defining the complex interplay between these proteins can further our understanding of the mechanisms controlling virus trafficking and evasion of the human immune system. In Chapter 3, the co-evolutionary analysis is applied to study possible domain-domain interactions of the mildew resistance gene o (MLO) protein family. A high degree of co-evolution between domains could identify potential domain interactions and/or domains that share a common binding partner. In Chapter 4, the co-

evolutionary an

proteins that int

possible protein-

were studied - t

their G-α subun

protein families,

their receptors.

proteins of uncha

were made.

evolutionary analysis is extended to quantitate the degree of co-evolution between proteins that interact. This approach allows for fast and objective identification of possible protein-protein interactions. Six systems of interacting protein families were studied - the syntaxin/unc-18 protein families, the adrenergic receptors and their G-α subunits, the TGF-β proteins and their receptors, the colicin/immunity protein families, the chemokines and their receptors, and the VEGF proteins and their receptors. From this analysis, inferences about the interaction partners for proteins of uncharacterized binding specificities in the TGF-β and syntaxin families were made.

UCSF LIBRARY

Table of Contents

UCSF LIBRARY

List of Tables

List of Figures

Cellular sign
to define and descr
how cells communi
In the post-genomi
growing knowledge
cellular machinery,

Traditionally
biochemical, and b
predicted proteins h
methods. These
systems (Fields &
and correlated mR
approaches such as
(Pellegrini et al., 1
disadvantages, the
protein interaction
proteins are suppo
combining ortholo
2002).

In this diss
evolutionary prin
understanding of p

Introduction

Cellular signaling is an essential component of biological processes. The ability to define and describe cellular signaling pathways can enable scientists to understand how cells communicate temporally and spatially in both normal and pathological states. In the post-genomic era, a vast amount of experimental data has contributed to the growing knowledge about signaling pathways. However, in order to fully understand the cellular machinery, all the interactions between the proteins need to be defined.

Traditionally, protein interactions have been studied individually using genetic, biochemical, and biophysical techniques. The large amount of newly discovered or predicted proteins has created a need for the use of high-throughput interaction-detection methods. These methods include experimental methods such as yeast two-hybrid systems (Fields & Song, 1989), mass spectrometry (Gavin et al., 2002; Ho et al., 2002), and correlated mRNA expression profiles (DeRisi et al., 1996), as well as computational approaches such as gene fusion (Marcotte et al., 1999) and phylogenetic profile analysis (Pellegrini et al., 1999). While each of these techniques has alternate advantages and disadvantages, these approaches have provided a wealth of valuable information about protein interactions. Currently, the most accurately annotated datasets of interacting proteins are supported by more than one of these techniques, demonstrating the utility of combining orthologous methods to produce more reliable interaction data (Tong et al., 2002).

In this dissertation, I present several computational approaches that integrate both evolutionary principles and experimental information in order to enhance our understanding of protein ligand interactions. The underlying concept in my approach is

based on the hypot
information that is
information. Based c
ancestor, evolution is
relationships among
protein's divergent e
experimental informa
interacting protein fa
of receptor-ligand int

In Chapter 1,
and their interaction
partner's binding su
(Atwell et al., 1997;
proteins that belong t
we compare the bindi
its ligand superfami
evolution between t
inferring binding
superfamily. This
analysis between t
terminal domain o
evolution between
used to demonstrat

based on the hypothesis that the evolutionary history of a protein family provides information that is implicitly orthogonal to experimental and other computational information. Based on the theory that all organisms are linked via descent to a common ancestor, evolution is an intrinsic theme in biological research. By studying the pattern of relationships among and between protein families, it is possible to associate trends in a protein's divergent evolution to its structure and function. Using previously determined experimental information, I was able to quantitate the degree of co-evolution between interacting protein families, and my results provide a strong base for the characterization of receptor-ligand interactions where the ligands are also proteins.

In Chapter 1, the underlying hypothesis of this dissertation is presented - proteins and their interaction partners must co-evolve so that any divergent changes in one partner's binding surface are complemented at the interface of its interaction partner (Atwell et al., 1997; Jespers et al., 1999; Moyle et al., 1994; Pazos et al., 1997). For proteins that belong to large superfamilies, the issue of co-evolution becomes apparent as we compare the binding specificity of a receptor superfamily to the binding specificity of its ligand superfamily. I have developed a method to quantitate the degree of co-evolution between two protein families that interact and used this as a measure for inferring binding specificity between the ligand superfamily and the receptor superfamily. This approach is applied to two different protein systems. Correlation analysis between the two phylogenetic distances of the N-terminal domain and C-terminal domain of phosphoglycerate kinase demonstrates the high degree of co-evolution between two interacting domains within a protein. This method was also used to demonstrate the high degree of co-evolution between the chemokines and their

receptors. This
significant role in

High corre
protein families ca
with uncharacteriz
that the closest sec
neighbors of its r
uncharacterized pr
proteins within the
the phylogenetic tre

Chapter 2
chemokine recepto
cytomegalovirus
Cytomegalovirus en
chemokine receptors
al., 1999). The CMV
bind to human ch
intracellular signalin
chemokines and ch
evasion. As more
uncharacterized CM
being discovered. B
receptors of both

receptors. This discovery was of particular importance since this system plays a significant role in a wide variety of human diseases.

High correlation scores between the calculated phylogenies of two interacting protein families can be an effective measure for assigning binding preferences to proteins with uncharacterized binding specificities. Since I could show for characterized proteins that the closest sequence neighbors of a ligand are far more likely to bind to the closest neighbors of its receptor, I was able to make inferences about binding partners for uncharacterized proteins. In Chapter 1, I also identified possible interacting pairs of proteins within the chemokine and chemokine receptor families by visual inspection of the phylogenetic trees.

Chapter 2 extends the co-evolutionary studies on mammalian chemokines and chemokine receptors to studying the interplay between human cellular and human cytomegalovirus (CMV)-encoded chemokines and chemokine receptors. Cytomegalovirus encodes genes that are similar in sequence to human chemokines and chemokine receptors (Chee et al., 1990; Gao et al., 1993; Neote et al., 1993; Penfold et al., 1999). The CMV-encoded chemokines and chemokine receptors have been shown to bind to human chemokines and chemokine receptors, thereby initiating various intracellular signaling processes. Chapter 2 discusses the role of these CMV-encoded chemokines and chemokine receptors in virus trafficking and human immune system evasion. As more genes are being cloned, there are a growing number of functionally uncharacterized CMV-encoded chemokine-like and chemokine receptor-like proteins being discovered. By analyzing the potential co-evolution of chemokines and chemokine receptors of both human and viral origin, we can augment our experimental

understanding of b
and their role in the

In Chapter
functional role of
transmembrane (T
suggest that it has
1993). However, i
defense processes.
that make up the M
A high level of co
bind to a common
intracellular and/or
function(s) should

Previously
protein families t
evolutionary analy
interact but also to
possible interacting
objective manner
systems. In order t
- the syntaxin/unc-
the TGF-β protein
chemokines and the

understanding of both human and viral chemokine and chemokine receptor interactions and their role in the mechanisms of viral infection and persistence.

In Chapter 3, I continue to apply the co-evolutionary approach to understand the functional role of mildew resistance gene o (Mlo), the only known family of seven transmembrane (TM) proteins in plants (Devoto et al., 1999). Mutant phenotypes of Mlo suggest that it has a role in cell death protection (Buschges et al., 1997; Wolter et al., 1993). However, it is not yet understood at the molecular level how Mlo modulates plant defense processes. I apply the co-evolutionary analysis to the fifteen separate domains that make up the Mlo protein in order to ascertain possible interactions between domains. A high level of correlation between domains can also indicate that these domains may bind to a common binding partner, thereby suggesting domains that could be involved in intracellular and/or extracellular signaling processes. Further characterization of Mlo function(s) should help to elucidate its role in plant defense and stress responses.

Previously I had shown that one could measure the co-evolution between two protein families that are known to interact. In Chapter 4, we extended the co-evolutionary analysis not only to quantitate the co-evolution between two families that interact but also to measure the co-evolution between single proteins in order to identify possible interacting proteins. This approach should provide scientists with a fast and objective manner for inferring binding partners within particular protein interaction systems. In order to test this hypothesis, I chose six different protein interaction systems - the syntaxin/unc-18 protein families, the adrenergic receptors and their G-α subunits, the TGF-β proteins and their receptors, the colicin/immunity protein families, the chemokines and their receptors, and the VEGF proteins and their receptors. My analysis

of these six systems r
employ in order to ma
physiology.

The large amo
challenge to scientist
proteins in cellular
computational approa
interacting protein fa
interacting proteins, a
protein systems. Ove
using evolutionary n
function.

References

Atwell, S., Ultsch, M.
remodeled pro

Buschges, R., Hollric
Daelen, R., va
Salamini, F. &
element of pla

Chee, M. S., Bankier,
Hutchison, C.

of these six systems revealed the various mechanisms of co-evolution binding proteins employ in order to maintain their binding interfaces and their functional role in cellular physiology.

The large amounts of data accumulated from the whole genome efforts present a challenge to scientists to understand the role of these newly discovered genes and proteins in cellular signaling processes. In the following chapters, I present a computational approach that quantitates the degree of co-evolution between two interacting protein families, extend this approach to quantitatively identify pairs of interacting proteins, and explore the validity of this method by applying it to various protein systems. Overall, this dissertation attempts to demonstrate the applicability of using evolutionary models to augment our experimental understanding of protein function.

References

Atwell, S., Ultsch, M., De Vos, A. M. & Wells, J. A. (1997). Structural plasticity in a remodeled protein-protein interface. *Science* 278, 1125-8.

Buschges, R., Hollricher, K., Panstruga, R., Simons, G., Wolter, M., Frijters, A., van Daelen, R., van der Lee, T., Diergaarde, P., Groenendijk, J., Topsch, S., Vos, P., Salamini, F. & Schulze-Lefert, P. (1997). The barley Mlo gene: a novel control element of plant pathogen resistance. *Cell* 88, 695-705.

Chee, M. S., Bankier, A. T., Beck, S., Bohni, R., Brown, C. M., Cerny, R., Horsnell, T., Hutchison, C. A., 3rd, Kouzarides, T., Martignetti, J. A. & et al. (1990). Analysis

of the protein-coc
AD169. *Curr To*
DeRisi, J., Penland, L.,
Su, Y. A. & Tre
expression patte
Devoto, A., Piffanelli,
Schulze-Lefert,
diversity of the
Fields, S. & Song, O.
interactions. N
Gao, J. L., Kuhns, D.
P. M. (1993).
inflammatory
Gavin, A. C., Bosch
Rick, J. M.,
Brajenovic,
Rudi, T., Gr
A., Copley.
Bouwmees
Furga, G. (
analysis of
Ho, Y., Gruhler,
Taylor, P.

of the protein-coding content of the sequence of human cytomegalovirus strain AD169. *Curr Top Microbiol Immunol* 154, 125-69.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A. & Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 14, 457-60.

Devoto, A., Piffanelli, P., Nilsson, I., Wallin, E., Panstruga, R., von Heijne, G. & Schulze-Lefert, P. (1999). Topology, subcellular localization, and sequence diversity of the Mlo family in plants. *J Biol Chem* 274, 34993-5004.

Fields, S. & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* 340, 245-6.

Gao, J. L., Kuhns, D. B., Tiffany, H. L., McDermott, D., Li, X., Francke, U. & Murphy, P. M. (1993). Structure and functional expression of the human macrophage inflammatory protein 1 alpha/RANTES receptor. *J Exp Med* 177, 1421-7.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. & Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141-7.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., S

Alfarano, C., De

Nielsen, P. A., R

Jespersen, H., P

B. D., Matthies

Durocher, D., N

Systematic ide

mass spectrom

Jespers, L., Lijnen, H.

De Maeyer, M

correlated mu

Molecular B

Marcotte, E. M., Pel

(1999). Dete

sequences. *S*

Moyle, W. R., Cam

(1994). Co-

Neote, K., DiGreg

cloning, fu

receptor. C

Pazos, F., Helmer-

mutations

Molecular

Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D. & Tyers, M. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 415, 180-3.

Jespers, L., Lijnen, H. R., Vanwetswinkel, S., Van Hoef, B., Brepoels, K., Collen, D. & De Maeyer, M. (1999). Guiding a docking mode by phage display: selection of correlated mutations at the staphylokinase-plasmin interface. *Journal Of Molecular Biology* 290, 471-9.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science* 285, 751-3.

Moyle, W. R., Campbell, R. K., Myers, R. V., Bernard, M. P., Han, Y. & Wang, X. (1994). Co-evolution of ligand-receptor pairs. *Nature* 368, 251-5.

Neote, K., DiGregorio, D., Mak, J. Y., Horuk, R. & Schall, T. J. (1993). Molecular cloning, functional expression, and signaling characteristics of a C-C chemokine receptor. *Cell* 72, 415-25.

Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *Journal Of Molecular Biology* 271, 511-23.

UCSF LIBRARY

Pellegrini, M., Marcot

Assigning prot

phylogenetic p

United States

Penfold, M. E., Dairag

W. & Schall, T

Proc Natl Aca

Tong, A. H., Drees, B

Evangelista, M

Hogue, C. W.,

experimental a

for peptide rec

Wolter, M., Hollriche

alleles to pow

defence mimi

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 96, 4285-8.

Penfold, M. E., Dairaghi, D. J., Duke, G. M., Saederup, N., Mocarski, E. S., Kemble, G. W. & Schall, T. J. (1999). Cytomegalovirus encodes a potent alpha chemokine. *Proc Natl Acad Sci U S A* 96, 9839-44.

Tong, A. H., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C. W., Fields, S., Boone, C. & Cesareni, G. (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 295, 321-4.

Wolter, M., Hollricher, K., Salamini, F. & Schulze-Lefert, P. (1993). The mlo resistance alleles to powdery mildew infection in barley trigger a developmentally controlled defence mimic phenotype. *Mol Gen Genet* 239, 122-8.

UCSF LIBRARY

Chapter 1

Co-Evolution of Proteins with their Interaction Partners

UCSF LIBRARY

This chapter was published as:

Goh CS, Bogan AA, Joachimiak M, Walther D and Cohen FE. (2000). Co-Evolution of Proteins with their Interaction Partners. *J Mol Biol*, **299**, 283-293.

Summary

The divergent evolution of proteins in cellular signaling pathways requires ligands and their receptors to co-evolve, creating new pathways when a new receptor is activated by a new ligand. However, information about the evolution of binding specificity in ligand-receptor systems is difficult to capture from sequences alone. We have used phosphoglycerate kinase (PGK), an enzyme that forms its active site between its two domains, to develop a standard for measuring the co-evolution of interacting proteins. The N-terminal and C-terminal domains of PGK form the active site at their interface and are covalently linked. Therefore, they must have co-evolved to preserve enzyme function. By building two phylogenetic trees from multiple sequence alignments of each of the two domains of PGK, we have calculated a correlation co-efficient for the two trees that quantifies the co-evolution of the two domains. The correlation coefficient for the trees of the two domains of PGK is 0.79, which establishes an upper bound for the co-evolution of a protein domain with its binding partner. The analysis is extended to ligands and their receptors, using the chemokines as a model. We show that the correlation between the chemokine ligand and receptor trees' distances is 0.57. The chemokine family of protein ligands and their G-protein coupled receptors (GPCRs) have co-evolved so that each subgroup of chemokine ligands has a matching subgroup of chemokine receptors. The matching subfamilies of ligands and their receptors create a framework within which the ligands of orphan chemokine receptors can be more easily determined. This approach can be applied to a variety of ligand and receptor systems.

UCSF LIBRARY

Keywords: co-evolution; protein interaction; ligand binding; G-protein coupled receptors; chemokines; orphan receptors

Introduction

The functions of proteins in biological systems are determined by the physical interactions that the proteins make with other molecules. Protein-protein binding is a subset of these interactions that is of primary importance in metabolic and signaling pathways. Proteins and their interaction partners must co-evolve so that any divergent changes in one partner's binding surface are complemented at the interface by their interaction partner (Atwell *et al.*, 1997; Jespers *et al.*, 1999; Moyle *et al.*, 1994; Pazos *et al.*, 1997). Otherwise, the interaction between the proteins is lost, along with its function. However, the co-evolution of interaction partners at the level of whole protein families is not well understood. Most of our current understanding of these interactions comes from genetic and biochemical experiments such as the common yeast two-hybrid assay (Fields & Song, 1989). Here, we consider if evolutionary information, in the form of statistical comparisons between the phylogenetic trees of protein families that interact with one another, can be used to recognize these interactions.

UCSF LIBRARY

Recent advancements in using sequence information from completed genomes have improved the ability to predict general groups of interaction partners in the absence of experimental data using computational techniques. Two of these methods rely on gene fusion events to predict likely interacting genes, based on the assumption that genes that become fused into a single gene in any organism are likely to interact in other organisms (Enright *et al.*, 1999; Marcotte *et al.*, 1999a). Another approach has been to compare the presence and absence of homologous genes across multiple genomes to infer the involvement of a particular gene in a pathway involving other genes with similar profiles across multiple genomes (Pellegrini *et al.*, 1999). A combined algorithm that

incorporates these appr
recently been published
broadly defining funct
building general pathw
correlated divergent ev
ligand-receptor signali

Ligand recepto
receptor, or conversel
evolution of a ligand
necessary to quantify
including the biologic
to interact functiona
correlation between t
a receptor family.
single gene was use
and their interaction

Co-evolution of Do

The co-evol
co-evolution of pro
single protein are
relationship betwe

incorporates these approaches, and also messenger RNA expression comparisons, has recently been published (Marcotte *et al.*, 1999b). These approaches are quite useful for broadly defining functions of uncharacterized genes in completed genomes and for building general pathway information. However, they are not optimized to analyze the correlated divergent evolution of proteins and their interaction partners within a single ligand-receptor signaling system.

Ligand receptor systems often have multiple ligands that interact with a single receptor, or conversely, many receptors for a single ligand. To understand the co-evolution of a ligand gene family with its corresponding receptor gene family, it is necessary to quantify the correlated divergent evolution of the two families while including the biologically relevant pairings between ligands and receptors that are known to interact functionally. We have developed a method to quantitatively measure the correlation between the phylogenetic tree of a ligand family with the phylogenetic tree of a receptor family. The co-evolution of two interacting protein domains fused into a single gene was used to establish a guideline for analyzing the co-evolution of proteins and their interaction partners.

UCSF LIBRARY

Co-evolution of Domains in a Single Protein

The co-evolution of domains within a single protein is better understood than the co-evolution of proteins that are produced from different genes. Since domains within a single protein are covalently linked to one another by the polypeptide chain, the relationship between any two domains that interact with one another is one to one. We have chosen phosphoglycerate kinase (PGK) as a model system for quantifying co-evolution.

PGK is a two domain protein with the enzyme active site formed by the interface between the two domains (Figure 1.1) (Banks *et al.*, 1979; Blake & Evans, 1974). PGK catalyzes the transfer of a phosphoryl-group from 1,3-bis-phosphoglycerate to ADP to form 3-phosphoglycerate and ATP, a critical step in glycolysis. A functional active site is achieved by the closing of the hinge between the two domains which positions the two substrates for the reaction (Bernstein *et al.*, 1997). Since the function of this enzyme depends on an active site that is formed between two independent domains, a working enzyme requires the two domains to have co-evolved. Any change in the N-terminal domain that perturbs the activity of the enzyme must be selected against, or subsequently compensated for, by a correlated change in the C-terminal domain. Because these two interacting domains are covalently linked, there is no ambiguity about each domain's interaction partner. For these reasons, PGK can be viewed as an example of co-evolution between two interacting domains. It is an ideal example for our statistical method of quantifying co-evolution between binding partners.

A multiple sequence alignment of PGKs from a vast array of species built with PSI-BLAST (Altschul *et al.*, 1997) was divided into two independent alignments, one for the N-terminal domain and another for the C-terminal domain (Figure 1.1). The short linking regions, which are not directly involved in forming the active site, were left out of the two domain alignments. As a result, two phylogenetic trees were generated based on the pairwise sequence distances in the alignments, one tree for each domain (Figure 1.2). To quantify the similarity of the two trees we calculated the linear correlation coefficient

UCSF LIBRARY

between the set of

equivalent distances i

between the domains

between the set of all pairwise distances in tree 1 (N-terminal domain) with the equivalent distances in tree 2 (C-terminal domain) based on the actual covalent linkages between the domains (see Methods).

UCSF LIBRARY

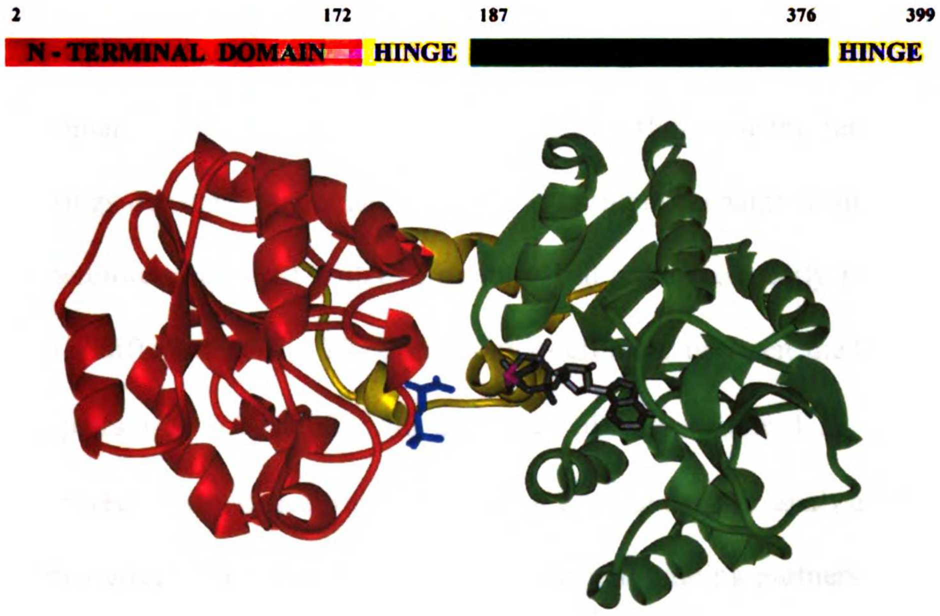


Figure 1.1. A ribbon diagram of the *T. maritima* phosphoglycerate kinase (PGK) structure (PDB 1vpe). The N-terminal domain (residues 2-172) is in red, the C-terminal domain (residues 187-376) is in green, and the hinge region (residues 173-186 and 377-399) is in yellow. The active PGK complex exhibits a hinge motion between the two terminal domains, bringing the two substrate ligands, 3-phosphoglycerate (blue) and ADP (grey) into close proximity (Bernstein et al., 1997). The functional active site is formed at the interface of the two domains.

UCSF LIBRARY

For the N and C-terminal domain trees, the correlation coefficient was 0.79±0.01, with a *z*-score of 41.91 (Table 1.1), indicating that the divergent evolution of the N-termini from one another is highly correlated to the divergent evolution of the C-termini from one another. To validate that this correlation was a meaningful measure of the co-evolution of the two domains, we recalculated the correlation coefficient using randomly chosen incorrect pairings between the domains. N and C-terminal domains from a single PGK gene were therefore not paired with one another, but were incorrectly matched with a domain from a different PGK. The correlation coefficient between the trees for these non-binding pairs was 0.00±0.02, with a *z*-score of 0.29 (Table 1.1). The lack of correlation between mismatched pairs serves as a control for our analysis method and shows that the correct linkage of domains with their real binding partners is required to observe co-evolution. To further control for the effects of speciation, as opposed to co-evolution, we also calculated the correlation coefficient between the tree for full length PGKs from 17 different species and a tree for topoisomerases (an enzyme that does not interact with PGK) from the same 17 species. The correlation coefficient for these two trees is 0.54±0.08 with a *z*-score of 6.25. This lower correlation coefficient suggests that, while speciation is an important effect, the higher correlation between the trees of the PGK N and C-terminal domains is due to co-evolution and not just speciation.

UCSF LIBRARY

Table 1.1 Correlation coefficients and related statistics

PGK N-Terminus and PGK C-Terminus

Binding Pairs:
Correlation Coefficient: 0.79±0.01
z-score: 41.91
P-value: 0.00

Non-Binding Pairs:
Correlation Coefficient: 0.00±0.02
z-score: 0.29
P-value: 0.77

Chemokines and Chemokine Receptors

Binding Pairs:
Correlation Coefficient: 0.57±0.02
z-score: 21.82
P-value: 0.00

Non-Binding Pairs:
Correlation Coefficient: 0.01±0.03
z-score: 0.41
P-value: 0.68

Human-only Chemokines and Chemokine Receptors

Binding Pairs:
Correlation Coefficient: 0.44±0.04
z-score: 11.23

PGKs and Topoisomerases:

Species Pairs:
Correlation Coefficient: 0.54±0.08
z-score: 6.25

Binding pairs refer to the pairs of interacting partners used in our statistical analysis (see Methods). They are either covalently linked (in the case of PGK's two domains) or experimentally known to bind one another (in the case of the chemokines and their receptors). Non-binding pairs were chosen at random and are not believed to interact. Since PGKs and topoisomerases do not bind to one another, pairings were done by species.

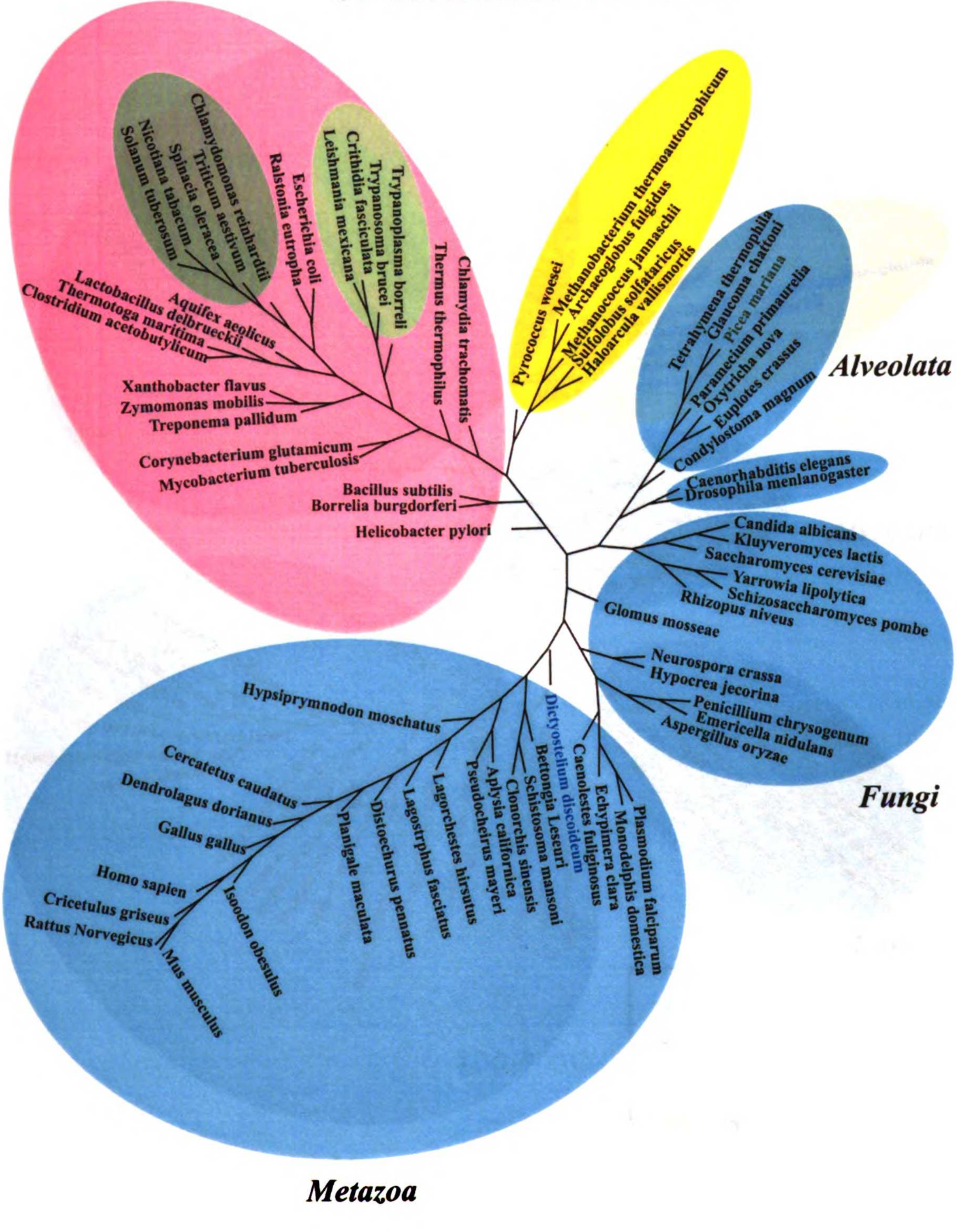
UCSF LIBRARY

The quantitative recognition of the co-evolution of the two domains of PGK was fully expected, since the two domains are linked to one another and must interact in order to function as an enzyme. However, a perfect correlation was not seen, since irregularities in the coordinated evolution of a single gene do occur, albeit relatively infrequently. For example, gene duplication or acquisition followed by domain swapping might allow for pairings of N and C-terminal domains that did not diverge together. It appears that this type of unexpected pairing of distantly related domains has occurred in the black spruce tree *Picea mariana*. Its PGK C-terminal domain clusters with those of other closely related viridiplantae whose PGKs appear to come from a eubacterial lineage (Figure 1.2b). However, the N-terminal domain of *Picea mariana* PGK is more similar to the eukaryotic alveolata than to the other viridiplantae N-termini, which remain with the eubacterial lineage (Figure 1.2a). The clustering of viridiplantae and euglenozoa PGKs within the eubacterial lineage (Figure 1.2, in green and pink) suggests that, in those groups of eukaryotes, PGK has most likely evolved from the genetic material of an organelle with eubacterial origins.

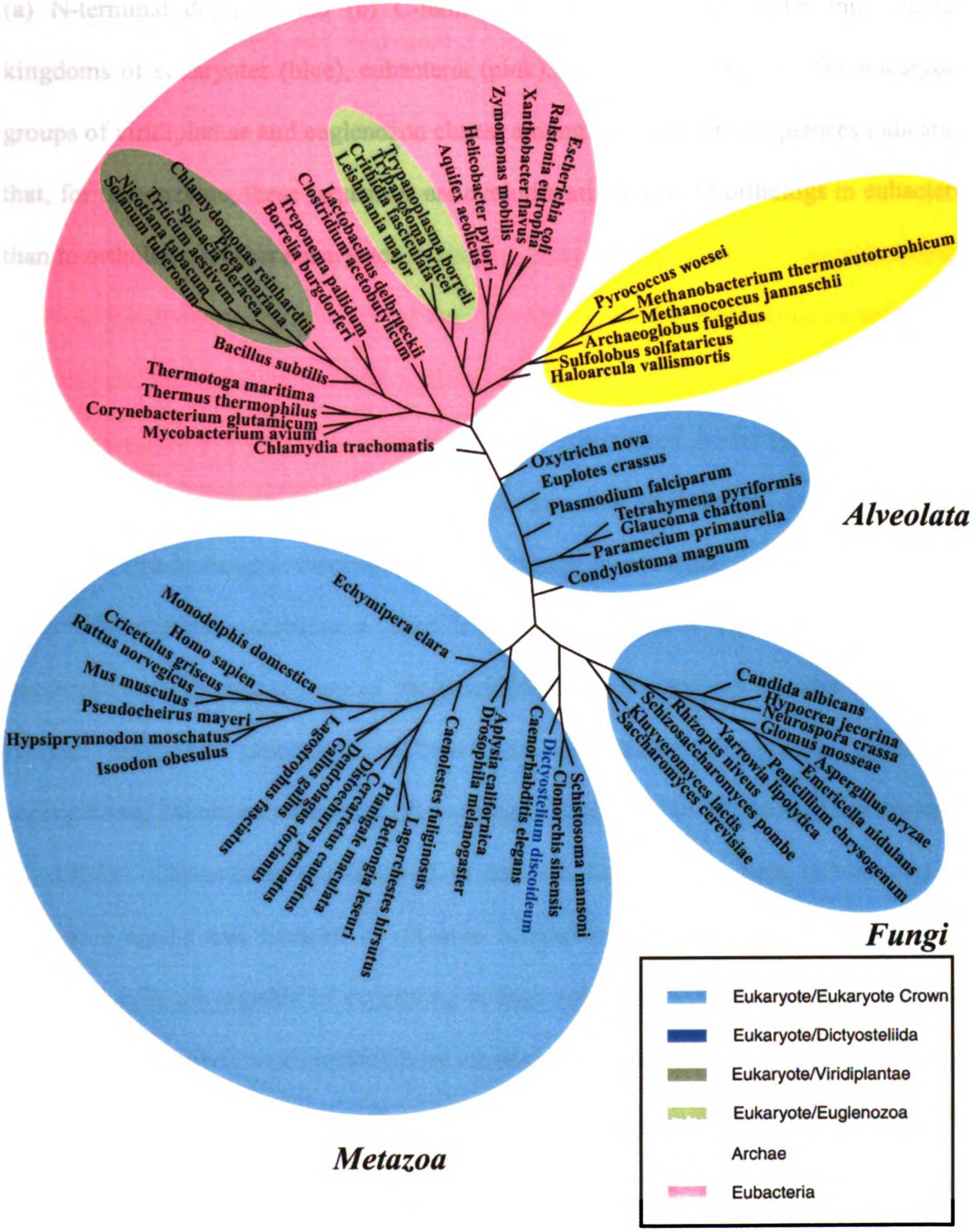
For a two domain protein like PGK, most of these domain swapping events are selected against, since function is rarely preserved. *Picea mariana* PGK is clearly an exception, not the rule. A few other organisms, such as *Drosophila melanogaster* and *Plasmodium falciparum*, show poor correlation between the two PGK domains in Figure 1.2, but the vast majority has clearly co-evolved. We conclude that a reasonable upper bound for a correlation coefficient in a system that has co-evolved is approximately 0.8. With this standard in mind from the PGK example, it is possible to evaluate the co-evolution of more complicated systems, such as ligands and their receptors.

UCSF LIBRARY

N-Terminus of PGK



C-Terminus of PGK



UCSF LIBRARY

Figure 1.2. The phylogenetic trees of the N-terminal and C-terminal domains of PGK. (a) N-terminal domains and (b) C-terminal domains of PGK cluster into separate kingdoms of eukaryotes (blue), eubacteria (pink), and archae (yellow). The eukaryotic groups of viridiplantae and euglenozoa cluster among the eubacteria sequences indicating that, for this enzyme, these sequences are evolutionarily closer to orthologs in eubacteria than to orthologs in other eukaryotes.

UCSF LIBRARY

Co-evolution of Ligands and Receptors

Ligands and receptors, like interacting domains, must co-evolve both to preserve necessary signaling pathways and to allow for the creation of new pathways during the evolution of an organism. However, it has been quite difficult to quantify or visualize the co-evolution of ligands and their receptors. We have applied our technique for measuring co-evolution to a ligand-receptor system that is well suited for this analysis, the chemokines and their transmembrane receptors. This is good model system for relating primary sequence knowledge to biological function. Our goal was to obtain information relevant to ligand-receptor binding specificity from sequence data.

Chemokines constitute a large family of *chemo*tactic cyto*kines* that activate transmembrane G-protein-coupled receptors (GPCRs) on the cell surface to regulate diverse biological processes. These processes include leukocyte trafficking, angiogenesis, hematopoiesis, and organogenesis (Baggiolini *et al.*, 1997; Oppenheim *et al.*, 1991). Chemokines are believed to be both beneficial in host defense against infectious agents and harmful in diseases marked by pathologic inflammation. All nucleated cells are capable of expressing at least some chemokines, and it appears that these molecules perform an extracellular messenger role in all tissues and systems of the body (Locati & Murphy, 1999). The chemokines are found in higher vertebrates and the ones included in this study are from various mammals (human, monkey, rat, mouse, pig, guinea pig, cow, sheep, dog, horse, rabbit, mangabey, gorilla, and chimpanzee), frog, and chicken.

UCSF LIBRARY

Recently, there has been increasing interest in chemokine receptors because CXCR4 and CCR5 have been found to be co-receptors for CD4-mediated HIV entry into cells (Premack & Schall, 1996). Not only do chemokines play a pivotal role in HIV infection, but they also exert other effects in inflammatory conditions and cancer (Wang *et al.*, 1998). Targeting specific chemokines and chemokine receptors may have therapeutic utility in inflammation, cancer, and infectious disease. The important role of chemokine signaling in disease, coupled with the wide variety of known chemokines and chemokine receptors, make this system ideal for studying the co-evolution of ligands and their receptors.

The chemokine nomenclature is defined by a cysteine signature motif where C is a cysteine and X is any amino acid residue (Clore & Gronenborn, 1995). They fall into four categories: CXC, CC, C, and CX_3C. Most of the known chemokines are members of the CXC or CC subfamilies. The C and the CX_3C chemokine subfamilies were discovered more recently. The first C chemokine found was lymphotactin; fractalkine was the first CX_3C chemokine discovered (Bazan *et al.*, 1997; Kelner *et al.*, 1994). We have selected various chemokine receptors and their cognate ligands for this analysis (Table 1.2).

UCSF LIBRARY

Table 1.2. Chemokines Receptors and Their Ligands

CC Chemokine Receptors	CC Chemokines
CCR1	MIP1α, RANTES, MCP3, HCC1, MPIF1, MIP5
CCR2	MCP1, MCP2, MCP3, MCP4, MCP5
CCR3	Eotaxin, MCP2, MCP3, MCP4, RANTES, Eotaxin2, MIP5
CCR4	TARC, MDC
CCR5	MIP1α, MIP1β, RANTES
CCR6	MIP3α
CCR7	MIP3β, SLC
CCR8	I-309, TARC, MIP1β
CCR9	TECK
CXC Chemokine Receptors	CXC Chemokines
CXCR1	IL-8
CXCR2	IL-8, GCP2, GRO-α, β, γ, ENA78, PGP
CXCR3	IP10, MIG
CXCR4	SDF1
CXCR5	BLC
C Chemokine Receptor	C Chemokine
XCR1	Lymphotactin
CX$_3$C Chemokine Receptor	CX$_3$C Chemokine
CX3CR1	Fractalkine

These experimentally determined binding partners (Baggiolini et al., 1997; Kim & Broxmeyer, 1999; Lu *et al.*, 1999; Rollins, 1997; Zaballos *et al.*, 1999) were used to calculate the correlation coefficient between the ligand and receptor trees (see Methods).

UCSF LIBRARY

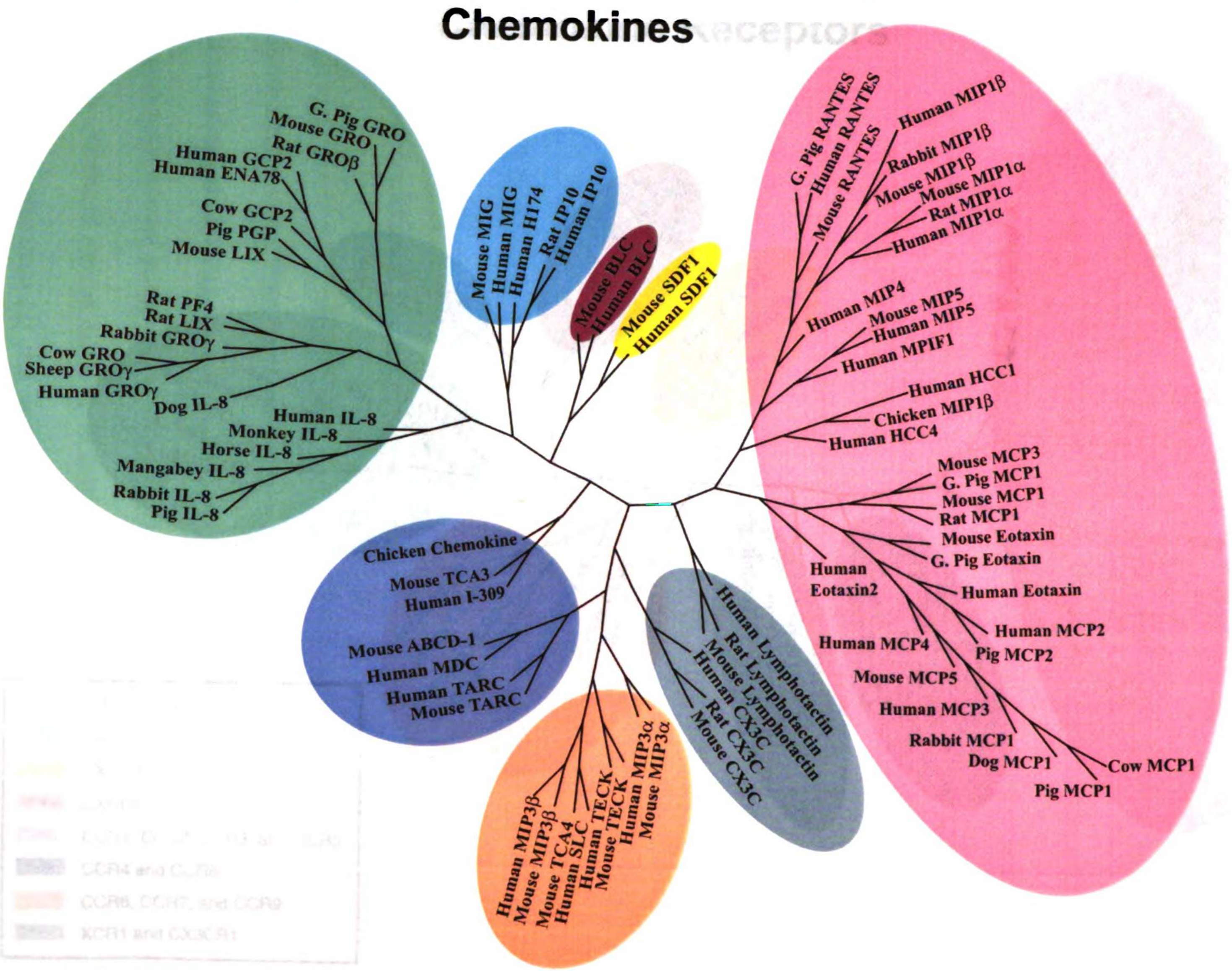
Our technique for mapping and quantifying the co-evolution of binding specificity was applied to the chemokine system. We built trees that show the correlated evolution of binding specificity for chemokines and their receptors (Figure 1.3). Using the known information regarding the binding of chemokines and their cognate receptors (Table 1.2) we calculated the correlation coefficient for the chemokine ligand and receptor trees. The correlation coefficient for these trees is 0.57±0.02 with a *z*-score of 21.82 (Table 1.1). Considering the upper bound of 0.8, which we have established using PGK, a two-domain system that has clearly co-evolved, the correlation coefficient of 0.57 indicates a very highly correlated co-evolution of the chemokines and their receptors. Since very few different (and less divergent) species were used in this case, the effects of speciation are much less significant for the chemokine system than they were for the PGK example. Still, we confirmed that speciation was not a major factor by calculating the correlation coefficient between the chemokines and their receptors within a single species. For only the human chemokines and their receptors, the correlation coefficient between the trees is 0.44±0.04 with a *z*-score of 11.23 and a *P*-value of 2.97×10^{-29}.

For any given chemokine, its closest sequence neighbors are far more likely to bind the closest neighbors of its receptor than to bind a randomly selected chemokine receptor. The analysis applies to all the chemokines in the phylogenetic tree (Figure 3) based on their known binding partners (Table 1.2). Our all-inclusive approach and calculation of a statistical correlation coefficient may explain why we find a high degree of co-evolution despite a previous study that concluded CC chemokines had not co-evolved closely with their receptors (Hughes & Yeager, 1999). Our control calculation

UCSF LIBRARY

was done based on incorrect binding partners chosen at random. For this random, non-binding map of ligands to receptors, the correlation coefficient was 0.01±0.03, with a *z*-score of 0.41 (Table 1.1). The non-correlation of randomly paired ligands and receptors demonstrates that the real biological interaction partners must be chosen to show co-evolution between ligands and their receptors. Since it is easy to add new sequences to phylogenetic trees, our approach creates a scalable framework allowing new chemokine or receptor sequences to be clustered based on their likely binding specificity. The search space for experimental determination of a novel family members' interaction partners is therefore greatly reduced. More detailed information about the binding specificity of the chemokines and their receptors can be obtained by analyzing the correlated phylogenetic trees (Figure 1.3).

UCSF LIBRARY


Chemokines
G. Pig GRO
Mouse GRO
Rat GROβ
Human GCP2
Human ENA78
Cow GCP2
Pig PGP
Mouse LIX
Rat PF4
Rat LIX
Rabbit GROγ
Cow GRO
Sheep GROγ
Human GROγ
Dog IL-8
Human IL-8
Monkey IL-8
Horse IL-8
Mangabey IL-8
Rabbit IL-8
Pig IL-8
Mouse MIG
Human MIG
Human H174
Rat IP10
Human IP10
Mouse BLC
Human BLC
Mouse SDF1
Human SDF1
G. Pig RANTES
Human RANTES
Mouse RANTES
Human MIP1β
Rabbit MIP1β
Mouse MIP1β
Mouse MIP1α
Rat MIP1α
Human MIP1α
Human MIP4
Mouse MIP5
Human MIP5
Human MPIF1
Human HCC1
Chicken MIP1β
Human HCC4
Mouse MCP3
G. Pig MCP1
Mouse MCP1
Rat MCP1
Mouse Eotaxin
G. Pig Eotaxin
Human Eotaxin2
Human Eotaxin
Human MCP2
Pig MCP2
Human MCP4
Mouse MCP5
Human MCP3
Rabbit MCP1
Dog MCP1
Cow MCP1
Pig MCP1
Chicken Chemokine
Mouse TCA3
Human I-309
Mouse ABCD-1
Human MDC
Human TARC
Mouse TARC
Human MIP3β
Mouse MIP3β
Mouse TCA4
Human SLC
Human TECK
Mouse TECK
Human MIP3α
Mouse MIP3α
Human Lymphotactin
Rat Lymphotactin
Mouse Lymphotactin
Human CX3C
Rat CX3C
Mouse CX3C

Chemokine Receptors

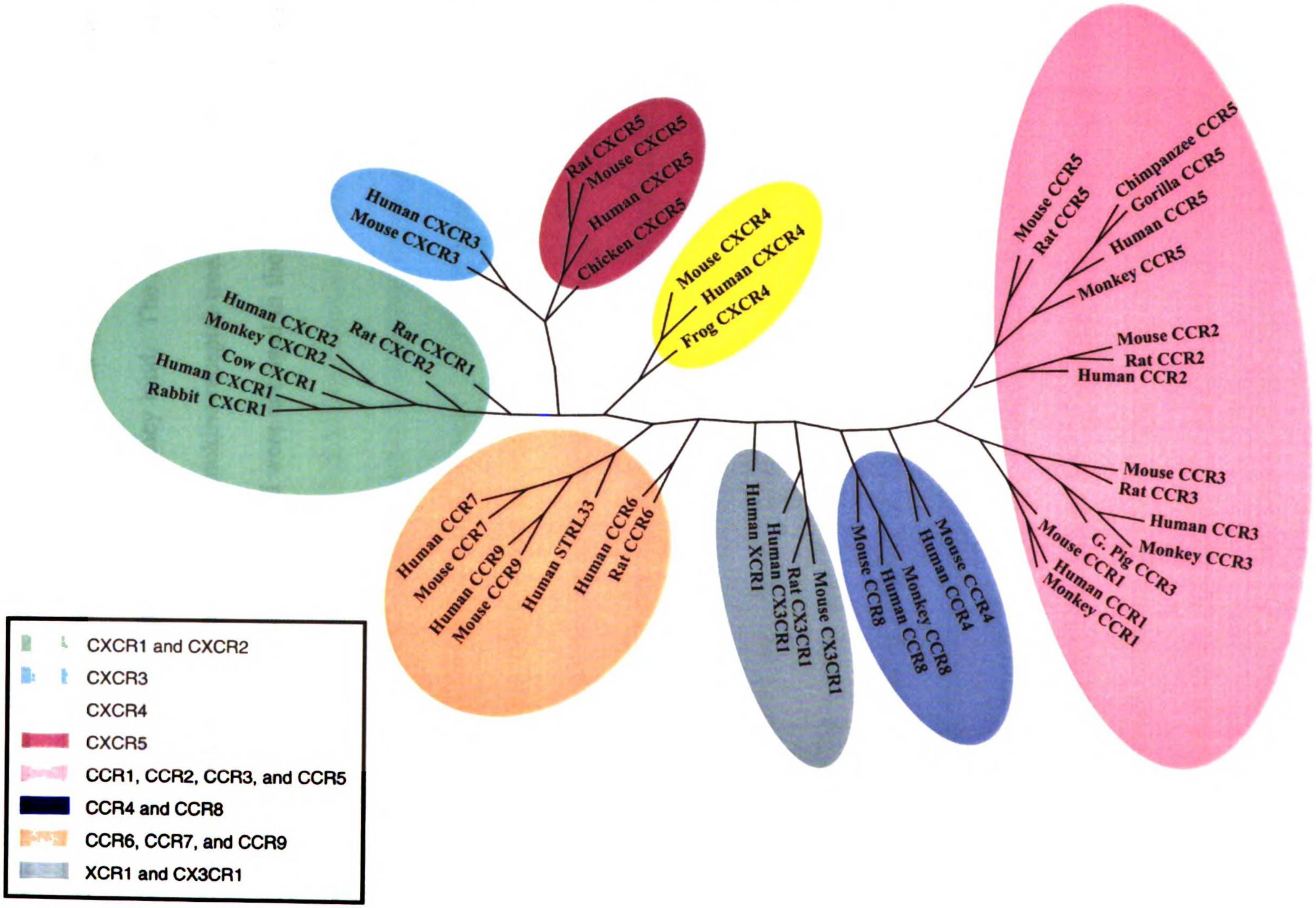


Figure 1.3. Phylogenetic trees of (a) chemokines and (b) chemokine receptors. The diagrams are colored by their clustering patterns to show similar groupings among the chemokines and the receptors to which they bind. The colored groups were chosen by eye based on the branching of the chemokine receptor tree. They are provided only as a guide for visualization of the data and were not used in the calculation of the correlation coefficients.

UCSF LIBRARY

Analysis of Chemokine Co-evolution

In Figure 1.3b, the CXC receptors cluster in a separate group from the CC receptors, with the C and CX_3C receptors forming their own group roughly equidistant from the CXC clusters and the main two groups of CC receptors. Among the CC receptors, CCR1, CCR2, CCR3, and CCR5 have sequences that are closely related to one another. CCR4 and CCR8 cluster together, as do CCR6, CCR7, CCR9, and the orphan receptor STRL33. This last subset of CC receptors falls as close to the CXC receptors as it does to the C and CX_3C receptors. Correspondingly, the ligands of the chemokine receptors form clusters that match the branches of the receptor tree (Figure 1.3a).

It is important to note, that there is some subjectivity in the assignments of clusters on the two trees (Figure 1.3). We have attempted to choose groupings that correspond to known physiological interactions wherever possible. For example, since CCR4 and CCR8 share a common ligand, TARC, we have chosen to group CCR4 and CCR8 together instead of grouping CCR8 with CX_3C1 (an equally plausible cluster based on the tree alone). However, these arbitrary choices were not used in the calculations of the correlation coefficients and therefore do not impact our statistical data.

The MIP chemokines (except MIP3) and RANTES group together, as do the nearby MCP chemokines and eotaxin (Figure 1.3a, colored pink). Subsets of these chemokines bind to CCR1, CCR2, CCR3, and CCR5 (Table 2), which form a cluster on the receptor tree (Figure 1.3b, also in pink). Similarly, MIP3α, MIP3β, TECK, and SLC cluster together (Figure 1.3a, in light red). MIP3α binds to CCR6; while MIP3β and SLC bind to CCR7. TECK binds to CCR9. The corresponding cluster can be found on

UCSF LIBRARY

he receptor tree where CCR6, CCR7, and CCR9 form a third subgroup of CC receptors long with the human orphan chemokine receptor STRL33 (Figure 1.3b, in light red).

Within the CXC chemokine receptors, CXCR1 and CXCR2 group together Figure 1.3b, in green). CXCR1 binds to IL-8; and CXCR2, with its broader specificity binds to IL-8, GCP2, the GROs, ENA78, and PGP. On the ligand tree, these chemokines also form a cluster within the other CXC chemokines (Figure 1.3a, in green). CXCR3, on its own branch of the CXC receptor cluster, binds to MIG and IP10 which cluster together on the chemokine tree (Figure 1.3, in blue). The human chemokine H174 also falls in this group. CXCR4 binds to SDF1 (Figure 1.3, in yellow) and CXCR5 binds to BLC (Figure 1.3, in magenta). The branching structure of the CXCR3-5 branches (Figure 1.3b in blue, magenta, and yellow) is not, however, identical to the branching structure of their ligands (Figure 1.3a in blue, magenta, and yellow). While the clusters still match between the trees, these differences in the branching patterns contribute to the imperfect correlation between the trees.

The grouping of the C and CX_3C chemokine receptors on the receptor tree corresponds with their ligands as well. The C chemokine, lymphotactin, and the CX_3C chemokine, fractalkine, can be grouped on the chemokine tree (Figure 1.3, in grey). This implies that the binding specificities of these two types of receptor are closer to one another than to CC or CXC receptors. However, because the trees were constructed using the neighbor-joining method and there is only one example of each of these two classes of chemokine receptors, there may be some bias toward pairing these sequences. Therefore, the C and CX_3C chemokines and their receptors may be less closely related than they appear on the trees.

UCSF LIBRARY

Since the chemokine and receptor trees cluster according to their binding specificities, we can begin to make inferences about possible ligands for orphan receptors and vice versa. (The "orphan" designation means that a cognate ligand or a cognate receptor is not known for a receptor or chemokine, respectively.) Several orphan chemokines and one orphan chemokine receptor were included in the trees (Figure 1.3). The orphan receptor STRL33 (Liao *et al.*, 1997) groups with CCR6 and CCR7. Based on the high correlation coefficient for our trees, we suggest that the orphan receptor STRL33 is likely to bind a chemokine that is from the corresponding group on the chemokine tree. This suggests that likely ligand candidates are chemokines (either known or not yet discovered) related to MIP3α, MIP3β, SLC, or TECK.

The human chemokine H174, which at the start of this study was an orphan, clusters with MIG and IP10 (Figure 1.3a, in blue), so we suggested that H174 binds a CXC chemokine receptor, most likely CXCR3 or one that is very similar in sequence. A recent independent experimental study has confirmed this prediction showing that H174 (also known as IP-9) is a high affinity ligand for CXCR3 (Tensen *et al.*, 1999). Two other orphan chemokines, HCC4 and MIP4 (Guan *et al.*, 1999; Hedrick *et al.*, 1998), cluster with their related CC chemokines (Figure 1.3a, in pink). We predict that the receptors of these orphan chemokines are likely to fall within the pink cluster of CCR receptors in Figure 1.3b.

PF4, another orphan chemokine, clusters with the ligands of CXCR1 and CXCR2. owever, it is known that PF4 does not bind CXCR1 or CXCR2 in its wild-type form. terestingly, engineered protein constructs containing a modification of the N-terminal quence of PF4 do bind to CXCR2 (Jones *et al.*, 1997). This implies that the sequence

USF LIBRARY

is competent for the predicted specificity, but its potential to interact has been suppressed by divergent evolution within specific regions of its N-terminus. In the case of PF4, the oligomerization state of the chemokine, may control its biological function. A recent study shows that tetrameric PF4 binds directly to glycosaminoglycans on the surface of neutrophils (Petersen *et al.*, 1999).

Conclusions

The co-evolution of the two domains of phosphoglycerate kinase was used to develop a guideline for quantifying co-evolution of proteins and their binding partners. Based on this guideline, the chemokines and their receptors were shown to have co-evolved. Our method was applied to orphan ligands and receptors in the search for orphans' binding partners. It provides a framework that significantly reduces the search space from all possible ligands or receptors to a small subset represented by a region of our phylogenetic tree. While the binding interactions of orphan ligands and receptors can only be proven experimentally, our analysis should aid in the rapid discovery of currently unknown chemokine signaling pathways.

UCSF LIBRARY

The approach is readily expandable to include new ligand and receptor sequences as they are discovered. It can also be applied to other systems of proteins and their interaction partners. Possible examples include other cytokines and kinases. It is also potentially useful for representing the evolution of ligand binding specificity in systems that have small molecule ligands, such as nuclear hormone receptors and other GPCRs once a suitable phylogeny of small molecules or the enzymes responsible for their biosynthesis can be established.

Methods

Sequence analysis

Sequences related to human CXCR1, IL-8, and phosphoglycerate kinase were retrieved using PSI-BLAST (Altschul et al., 1997) with default parameters and the complete non-redundant database. Multiple sequence alignments of the chemokine receptors, the chemokines, and the phosphoglycerate kinases were constructed based directly on the PSI-BLAST alignments. The multiple sequence alignment for PGK was divided into two alignments, one for each domain. The N-terminal domain alignment included amino acids 2-172 and the C-terminal domain included amino acids 187-376. Topoisomerase I sequences from 17 different species (including eukaryotes, eubacteria, and archae) were selected from the SWISSPROT database and aligned using ClustalW. The ClustalW phylogeny program was used to calculate a distance matrix by percent sequence divergence and to generate the trees with the neighbor-joining method (Saitou & Nei, 1987). The unrooted trees were drawn using the DrawTree program in PHYLIP (Felsenstein, 1993).

Correlation analysis

Distance matrices were generated from the multiple alignments using ClustalW (Thompson *et al.*, 1994). In order to quantify the co-evolution of interaction partners, we employed a linear regression analysis measuring the correlation between pairwise

UCSF LIBRARY

evolutionary distances among all proteins in a multiple sequence alignment. These were correlated with the evolutionary distances among the corresponding binding partners (or, in the case of PGK and topoisomerase I, the corresponding species, since these proteins do not bind). We defined X as a two-dimensional matrix of evolutionary distances in the receptor family (X was constructed as a NxN matrix, where N is equal to the number of receptors). For the corresponding ligands, a similar distance matrix, Y, was constructed. X_{ij} is the pairwise distance between sequence m_i and sequence m_j, and Y_{ij} signifies the pairwise distance between sequence n_i and sequence n_j (where n_i is experimentally known to bind to m_i and n_j is known to bind to m_j). In order to represent multiple ligands that bind to a single receptor, or vice versa, there were instances where the same ligand or receptor was represented multiple times in the matrix. Therefore in the cases where one ligand was known to experimentally bind to two different receptors, the ligand was represented as both n_i and n_j in matrix Y corresponding to the two different receptors, m_i and m_j, in matrix X. The correlation coefficient was then calculated for all the pairwise distances in matrix X and their corresponding distances in matrix Y.

We computed the linear correlation coefficient r (Pearson's correlation coefficient, (Press *et al.*, 1988)) defined as:

$$r = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(X_{ij}-\overline{X}\right)\left(Y_{ij}-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(X_{ij}-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(Y_{ij}-\overline{Y}\right)^2}}$$

with $-1 \le r \le +1$ where $\overline{X}$ is the mean over all X_{ij}'s, and $\overline{Y}$ is the mean over all Y_{ij}'s. In our context, X_{ij} and Y_{ij} are pairwise sequence similarity distances between N-terminal and C-terminal domains of PGK, or between chemokine receptors and their corresponding

UCSF LIBRARY

chemokines, respectively. Positive values of r would indicate a positive co-evolution; *i.e.* receptors that appear to be evolutionarily close, have ligands that, in turn, are more closely related than other pairs of any two ligands. By contrast, r-values of around zero would indicate no correlation, and negative values of r would indicate anti-correlation.

Estimation of statistical significance of correlation

The significance of the computed value r was assessed by a bootstrapping analysis yielding an estimate of the standard deviation of r given the size of our data set (Efron, 1979), and by an estimation of the probability of obtaining the observed value of r by chance (P-value). In the bootstrap analysis, we generated 1000 sets containing N pairwise distances randomly drawn (with replacement) from the N pairwise distances in the original set. For every such set we computed the bootstrap correlation coefficient r_b. The bootstrap interval; *i.e.* the interval of r_b accounting for 68% of the obtained values of r_b was obtained from the 16% (a) and 84% (b) percentiles in the histogram of the 1000 values r_b and the mean value of r_b from the 50% percentile. The bootstrap estimate of the standard deviation of the observed correlation then calculates as $\sigma_b = \dfrac{b-a}{2}$.

The P-value; *i.e.* the probability that the particular correlation coefficient r quantifying the co-evolution between chemokines and their receptors was obtained by chance, was obtained by randomly shuffling the pairwise distances between ligands and receptors. Thus the assignments of correspondence (ligand l_1 binds to receptor R_{l1}, and ligand l_2 binds to receptor R_{l2}) were replaced by random assignments and the correlation coefficient was computed as explained above. This process was repeated 1000 times. From the resulting 1000 values r_{rand} a z-score for the actual observed value r was

UCSF LIBRARY

calculated as $z = \frac{r - \bar{r}_{rand}}{\sigma_{rand}}$ where σ is the standard deviation of r_{rand} and $\bar{r}_{rand}$ is the mean (effectively zero for truly random data). The *P*-value is then obtained from $P = erfc(|z|)/\sqrt{2}$ where *erfc* is the complement error function.

Acknowledgements

A.A.B. was supported by a National Defense Science and Engineering Graduate Fellowship from the United States Department of Defense and by the Lloyd M. Kozloff Fellowship. This work was supported by grants from the NIH (to F.E.C.) We thank Jonathan Blake and John-Marc Chandonia for helpful discussions.

UCSF LIBRARY

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**(17), 3389-402.

Atwell, S., Ultsch, M., De Vos, A. M. & Wells, J. A. (1997). Structural plasticity in a remodeled protein-protein interface. *Science* **278**(5340), 1125-8.

Baggiolini, M., Dewald, B. & Moser, B. (1997). Human chemokines: an update. *Annual Review Of Immunology* **15**, 675-705.

Banks, R. D., Blake, C. C., Evans, P. R., Haser, R., Rice, D. W., Hardy, G. W., Merrett, M. & Phillips, A. W. (1979). Sequence, structure and activity of phosphoglycerate kinase: a possible hinge-bending enzyme. *Nature* **279**(5716), 773-7.

Bazan, J. F., Bacon, K. B., Hardiman, G., Wang, W., Soo, K., Rossi, D., Greaves, D. R., Zlotnik, A. & Schall, T. J. (1997). A new class of membrane-bound chemokine with a CX3C motif. *Nature* **385**(6617), 640-4.

Bernstein, B. E., Michels, P. A. & Hol, W. G. (1997). Synergistic effects of substrate-induced conformational changes in phosphoglycerate kinase activation [see comments]. *Nature* **385**(6613), 275-8.

Blake, C. C. & Evans, P. R. (1974). Structure of horse muscle phosphoglycerate kinase. Some results on the chain conformation, substrate binding and evolution of the molecule from a 3 angstrom Fourier map. *Journal Of Molecular Biology* **84**(4), 585-601.

Clore, G. M. & Gronenborn, A. M. (1995). Three-dimensional structures of alpha and beta chemokines. *Faseb Journal* **9**(1), 57-62.

Efron, B. (1979). Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review* **21**, 460-480.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90.

UCSF LIBRARY

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) 3.5c edit. Department of Genetics, University of Washington, Seattle.

Fields, S. & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* **340**(6230), 245-6.

Guan, P., Burghes, A. H., Cunningham, A., Lira, P., Brissette, W. H., Neote, K. & McColl, S. R. (1999). Genomic organization and biological characterization of the novel human CC chemokine DC-CK-1/PARC/MIP-4/SCYA18. *Genomics* **56**(3), 296-302.

Hedrick, J. A., Helms, A., Vicari, A. & Zlotnik, A. (1998). Characterization of a novel CC chemokine, HCC-4, whose expression is increased by interleukin-10. *Blood* **91**(11), 4242-7.

Hughes, A. L. & Yeager, M. (1999). Coevolution of the mammalian chemokines and their receptors. *Immunogenetics* **49**(2), 115-24.

Jespers, L., Lijnen, H. R., Vanwetswinkel, S., Van Hoef, B., Brepoels, K., Collen, D. & De Maeyer, M. (1999). Guiding a docking mode by phage display: selection of correlated mutations at the staphylokinase-plasmin interface. *Journal Of Molecular Biology* **290**(2), 471-9.

Jones, S. A., Dewald, B., Clark-Lewis, I. & Baggiolini, M. (1997). Chemokine antagonists that discriminate between interleukin-8 receptors. Selective blockers of CXCR2. *Journal Of Biological Chemistry* **272**(26), 16166-9.

Kelner, G. S., Kennedy, J., Bacon, K. B., Kleyensteuber, S., Largaespada, D. A., Jenkins, N. A., Copeland, N. G., Bazan, J. F., Moore, K. W., Schall, T. J. & al, e. (1994). Lymphotactin: a cytokine that represents a new class of chemokine. *Science* **266**(5189), 1395-9.

Kim, C. H. & Broxmeyer, H. E. (1999). Chemokines: signal lamps for trafficking of T and B cells for development and effector function. *Journal Of Leukocyte Biology* **65**(1), 6-15.

Liao, F., Alkhatib, G., Peden, K. W., Sharma, G., Berger, E. A. & Farber, J. M. (1997). STRL33, A novel chemokine receptor-like protein, functions as a fusion cofactor for both macrophage-tropic and T cell line-tropic HIV-1. *Journal Of Experimental Medicine* **185**(11), 2015-23.

Locati, M. & Murphy, P. M. (1999). Chemokines and chemokine receptors: biology and clinical relevance in inflammation and AIDS. *Annual Review Of Medicine* **50**, 425-40.

UCSF LIBRARY

Lu, B., Humbles, A., Bota, D., Gerard, C., Moser, B., Soler, D., Luster, A. D. & Gerard, N. P. (1999). Structure and function of the murine chemokine receptor CXCR3. *Eur. J. Immunol.* **29**, 3804-3812.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999a). Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**(5428), 751-3.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-86.

Moyle, W. R., Campbell, R. K., Myers, R. V., Bernard, M. P., Han, Y. & Wang, X. (1994). Co-evolution of ligand-receptor pairs. *Nature* **368**(6468), 251-5.

Oppenheim, J. J., Zachariae, C. O., Mukaida, N. & Matsushima, K. (1991). Properties of the novel proinflammatory supergene intercrine cytokine family. *Annual Review Of Immunology* **9**, 617-48.

Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *Journal Of Molecular Biology* **271**(4), 511-23.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **96**(8), 4285-8.

Petersen, F., Brandt, E., Lindahl, U. & Spillmann, D. (1999). Characterization of a neutrophil cell surface glycosaminoglycan that mediates binding of platelet factor 4. *Journal Of Biological Chemistry* **274**(18), 12376-82.

Premack, B. A. & Schall, T. J. (1996). Chemokine receptors: gateways to inflammation and infection. *Nature Medicine* **2**(11), 1174-8.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1988). *Numerical Recipes in C*, Cambridge University Press, Cambridge.

Rollins, B. J. (1997). Chemokines. *Blood* **90**(3), 909-28.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology And Evolution* **4**(4), 406-25.

UCSF LIBRARY

Tensen, C. P., Flier, J., Van Der Raaij-Helmer, E. M., Sampat-Sardjoepersad, S., Van Der Schors, R. C., Leurs, R., Scheper, R. J., Boorsma, D. M. & Willemze, R. (1999). Human IP-9: A keratinocyte-derived high affinity CXC-chemokine ligand for the IP-10/Mig receptor (CXCR3). *Journal Of Investigative Dermatology* **112**(5), 716-22.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**(22), 4673-80.

Wang, J. M., Deng, X., Gong, W. & Su, S. (1998). Chemokines and their role in tumor growth and metastasis. *Journal Of Immunological Methods* **220**(1-2), 1-17.

Zaballos, A., Gutiérrez, J., Varona, R., Ardavín, C. & Márquez, G. (1999). Cutting edge: identification of the orphan chemokine receptor GPR-9-6 as CCR9, the receptor for the chemokine TECK. *Journal Of Immunology* **162**(10), 5671-5.

UCSF LIBRARY

C

This

Rece
the I

Chapter 2

Viral Chemokine Receptors and Chemokines in Human Cytomegalovirus Trafficking and Interaction with the Immune System

UCSF LIBRARY

This chapter was published as:

Beisser PS, Goh CS, Cohen FE, and Michelson S. (2002). Viral Chemokine Receptors and Chemokines in Human Cytomegalovirus Trafficking and Interaction with the Immune System. *Curr Top Microbiol Immunol*, **269**, 203-234.

LIBRARY
LIBRAR
LIBRA

UCSF LIBRARY

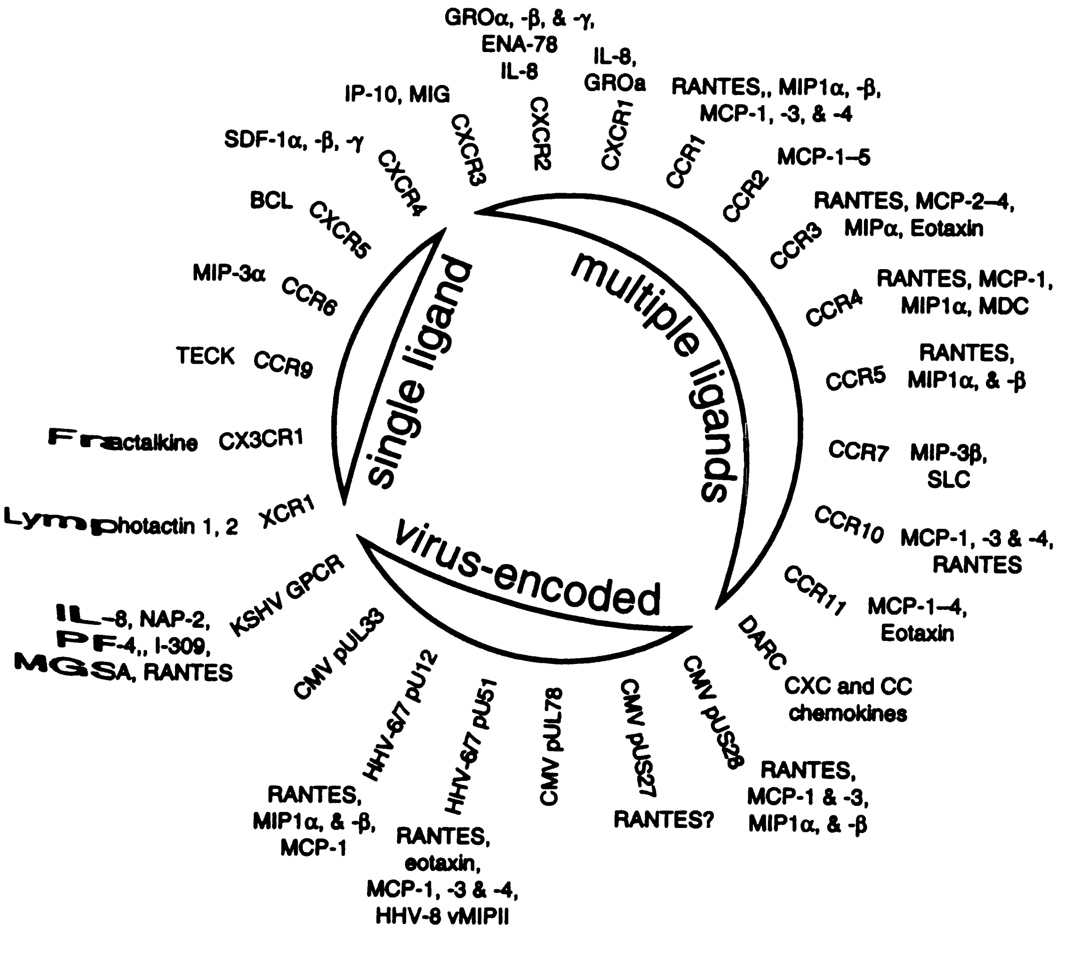
1 Introduction

Human CMV has devised numerous means of getting around detection by the immune system. Many CMV-specific genes encode molecules that interfere with both innate and adaptive immunity (see other chapters in this book). Some of these genes encode proteins that target antigen presentation, while others encode cytokines and chemokines, or cytokine or chemokine receptors. This review will focus on the human CMV homologs of chemokine receptors and induction of chemokines by CMV, and will discuss some of the potentials that these molecules have in virus trafficking during CMV infection and immune evasion.

2 Chemokines and Chemokine Receptors Interaction and Signaling

Chemokines are soluble mediators implicated in infiltration, inflammation and activation of leukocyte effector mechanisms. Many recent reviews have appeared, amongst which those that cover the new nomenclature of chemokines and their receptors (MURPHY et al. 2000), co-evolution of chemokine receptors and their ligands (GOH et al. 2000), chemokine-based lymphocyte trafficking (LOETSCHER et al. 2000), and viral anti-chemokines (MURPHY 2000). All chemokines have very similar overall structures, being composed of 3 beta sheets and an alpha helix, which separate the short N-terminal and the C-terminal domains. Chemokines are subdivided into 4 families based on the number and spacing of conserved cysteines: CXC with one amino acid (aa) separating the first 2 cysteines, CC with no intervening aa, CX3C with 3 intervening aa, and C with only one Cys residue. CXC chemokines can be further subdivided into "ERL^{+}", which are angiogenic, and ELR^{-}, which are usually angiostatic. Generally, CXC chemokines attract neutrophils and lymphocytes, whereas CC chemokines attract monocytes and macrophages (BAGGIOLINI et al. 1997). Almost all chemokines fall into either the CXC or

UCSF LIBRARY



UCSF LIBRARY

Figure 2.1. Human cellular and human CMV-encoded chemokine receptors and their corresponding ligands. Cellular chemokine receptors, which bind only one ligand (single ligand) or those that bind several ligands (multiple ligands), are shown. The human herpesvirus chemokine receptors (virus-encoded) and their ligands are given where determined. The old nomenclature has been used here. For the new nomenclature see Murphy et al (MURPHY et al. 2000).

UCSF LIBRARY

the
one
Cel
che
Sev
bin
bee
(rev
tho
tran
The
che
rec
N-t
imp
an
The
(Gi
int
bo
me
in
dif
en
tis
of

the CC families, since only two chemokines have been described for the C family and one for the CX3C family.

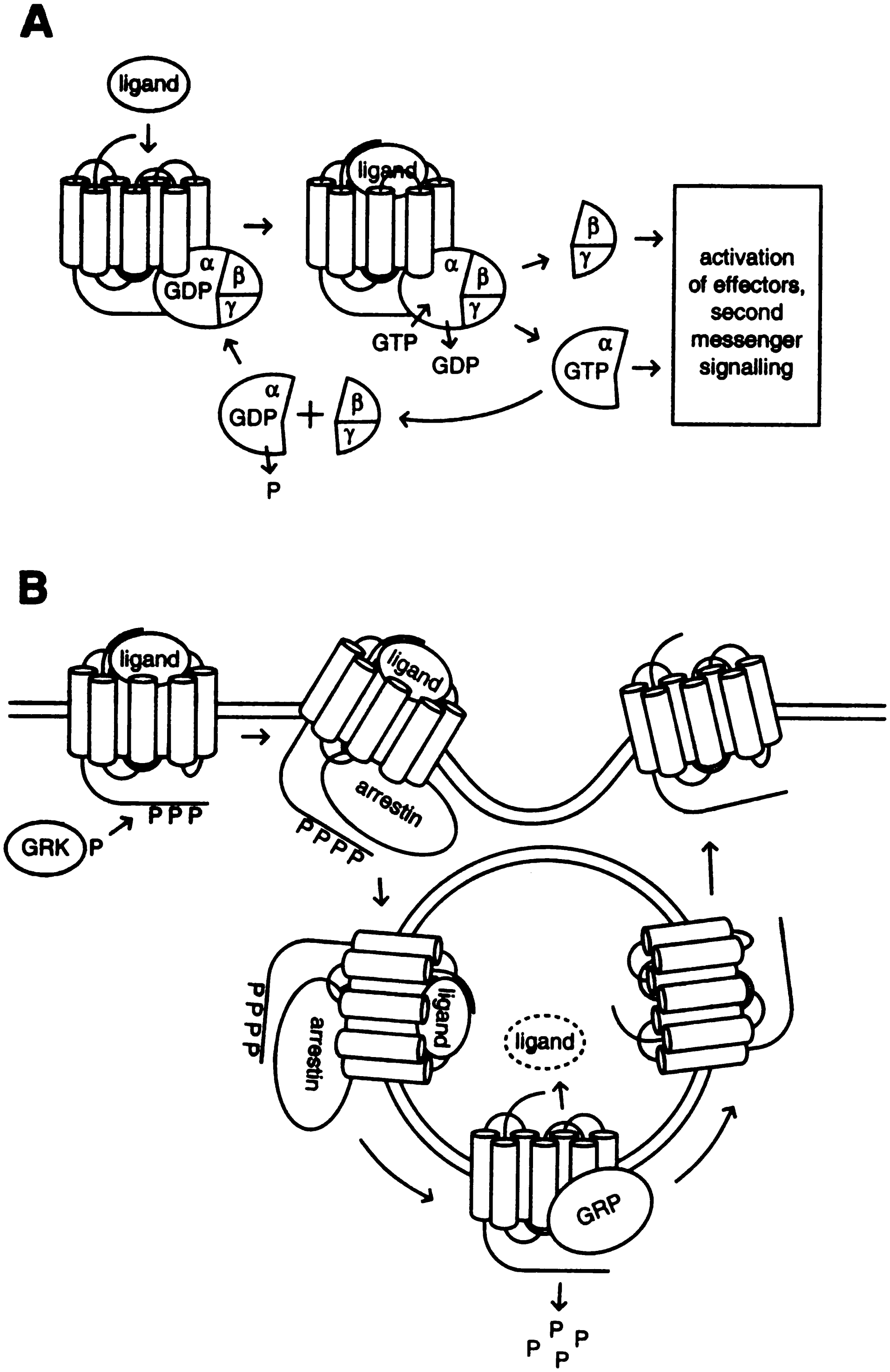
Cellular CXC chemokines bind only to CXC receptors (designated CXCR1--5) and CC chemokines bind only CCRs (designated CCR1--11) (review: (Murphy et al. 2000)). Several chemokines may bind to a given receptor and, conversely, several receptors may bind the same chemokine (Figure 2.1). For some CCRs and CXCRs, only one ligand has been found so far. These are often involved in homeostasis. Finally, Duffy antigen (reviewed in (Murphy et al. 2000)), found on erthyrocytes and endothelial cells, is thought to act as a "chemokine sink" and binds both CXC and CC chemokines, but transmits no intracellular signal.

The structure, as well as the mechanism, of ligand binding and signal transduction by chemokine receptors is similar to those of other members of the G protein-coupled receptor (GPCR) family (Selbie Hill 1998). Chemokine receptors have an extracellular N-terminal tail, 7 transmembrane domains, and 3 extracellular loops, which are all important for chemokine binding. In addition, the receptors have 3 intracellular loops and an intracellular C-terminal tail, which are essential for G protein binding and activation.

The mechanism of GPCR-mediated signaling is summarized in Fig. 2.2A (reviewed by (Gutkind 1998; Hamm 1998)). Chemokine-chemokine receptor complexes can be internalized through clathrin-dependent receptor endocytosis into endosomes, where the bound chemokine is released and degraded, and the receptor rerouted to the plasma membrane (Fig. 2.2B) (reviewed in (Signoret Marsh)). Receptor stimulation (reviewed in (Baggiolini 1998)) eventually leads to (i) differentiation, or inhibition of differentiation of leukocyte progenitors, (ii) rolling and attachment to blood vessel endothelial cells, as well as transendothelial migration, (iii) chemotaxis to inflamed tissue, or inhibition of chemotaxis in non-inflamed tissue, and (iv) induction of a variety of immunological responses such as cytotoxicity.

UCSF LIBRARY

UC
Francis
LIBRARY
UC
Franc
LIBRAR
UC
Fra
LIBRA



UCSF LIBRARY

Fig

cyc

pla

cha

repl

whi

case

Ca^{2}

the

retu

(B)

desc

long

thro

(GR

be i

the

GP

Figure 2.2. G protein-coupled receptor- (GPCR-) mediated signalling. (A) The GTP cycle in GPCR-mediated signalling. Note that the G proteins, as well as the effectors, are plasma membrane-bound. Upon ligand binding, the GPCR undergoes a conformational change that enables it to interact with G proteins -- the G protein-associated GDP is replaced by GTP, causing the G protein to dissociate in a G_α and a $G_{\beta\gamma}$ subunit, each of which are released from the receptor. Both G_α and $G_{\beta\gamma}$ subunits can activate signalling cascades that can result in the release of either inositol-3-phosphate (IP_3), cyclic-AMP or Ca^{2+}. G_α subunits have an intrinsic hydrolysis activity, resulting in dephosphorylation of the G_α-associated GTP. Upon dephosphorylation, G_α can reassociate with $G_{\beta\gamma}$, thereby returning to an inactivated state.

(B) Internalization of desensitized chemokine receptors. Chemokine receptors can be desensitized after an initial round of signalling -- i.e. modified such that they can no longer be activated through successive chemokine binding events. This is established through phosphorylation of the intracellular C-terminus of the receptor by GPCR kinases (GRK) and subsequent binding with β-arrestins. Upon desensitization, the receptor can be internalized by inclusion into clathrin-coated endosome vesicles. Upon internalization, the chemokine is release and degraded, whereas the receptor is dephosphorylated by GPCR phosphatase and rerouted to the plasma membrane.

UCSF LIBRARY

UC San Francisco LIBRARY

Tab

G

i

q

s

12

ki

te

Table 2.1. Activation Activities of G-Proteins

G_α subtypes (20)	Type of signal transduction	$G_{\beta\gamma}$subunits (6_β, 12_γ)[a]
i[c]	Inhibition of Adenylyl Cyclases and activation	--
	of PI3K[b]	+
	Activation of ion channels	
q	Activation of GRK[b]	+
	Activation of tyrosine kinases	+
	Activation of PLCβ[b]	+
s	Activation of Adenylyl Cyclases	--
12	Activation of ion channels	+
	Increase SAPK/JNK[b] activity	--
	rho-dependent induction of actin	--
	polymerization	--
	Induction of NO synthase	

[a] Numbers in parentheses denote the number of family members.

[b] Abbreviations: PI3K, phosphoinositol-3-kinase; GRK, G protein-coupled receptor kinase; PLCβ, phospholipase C-β; SAPK, stress activated protein kinase; JNK, c-Jun N-terminal kinase.

[c] $G_{\alpha i}$ proteins are sensitive to inhibition by Pertussis toxin.

MSU LIBRARY

3 CMV-Encoded Chemokines

CMV produces a functional chemokine, encoded by UL146 and designated vCXC-1 (PENFOLD et al. 1999). This gene was found in the genome of the Toledo strain of CMV (CHA et al. 1996). UL146 encodes an ERL$^+$, CXC-type chemokine, which, like IL-8, probably does not bind to CMV GPCRs US28 or US27. The vCXC-1 chemokine attracts human peripheral blood neutrophils. It binds with high affinity to CXCR2-transfected, but not to CXCR1-transfected mouse fibroblasts, as well as to freshly isolated human neutrophils. In addition, the downstream ORF UL147 also shows homology to CXC chemokines, but lacks an ERL motif. The Towne strain of CMV carries a UL146-like gene (UL152) which is in the opposite orientation to that of UL146. Whether UL147 and UL152 encode functional chemokines remains to be investigated. A detailed review is given by Dr. E. Mocarski in chapter 14.

4 CMV-Encoded Chemokine Receptors

Potential human CMV chemokine receptors were first discovered when Chee et al. (CHEE et al. 1990a) sequenced the genome of strain AD169. They described 3 receptor homologues designated UL33, US28 and US27 according to their genomic locations within the long unique (UL) and short unique (US) regions of the genome (CHEE et al. 1990b). Subsequently, Gompels et al. (GOMPELS et al. 1995) defined another potential GPCR gene by homology of a GPCR gene found in the Herpesvirus 6 genome with CMV UL78. CMVs from non-primates carry positional and sequence equivalents of the UL33 and UL78 genes. However, it is important to note that so far only human CMV carries the GPCR genes US27 and US28, thus restricting *in vivo* study of the latter receptors.

In the following sections, the sequence, transcription and expression properties of the CMV chemokine receptor genes and their gene products will be outlined, as will be their known and anticipated chemokine binding capacities.

UC Francisco
LIBRARY
LIBRARY
UC Francisco
LIBRARY

4.1 S

The f

cloni

(Nec

sequ

ident

foun

the p

of 3

othe

Clo

appr

tran

US2

tha

end

not

wit

has

pos

(Fi

co

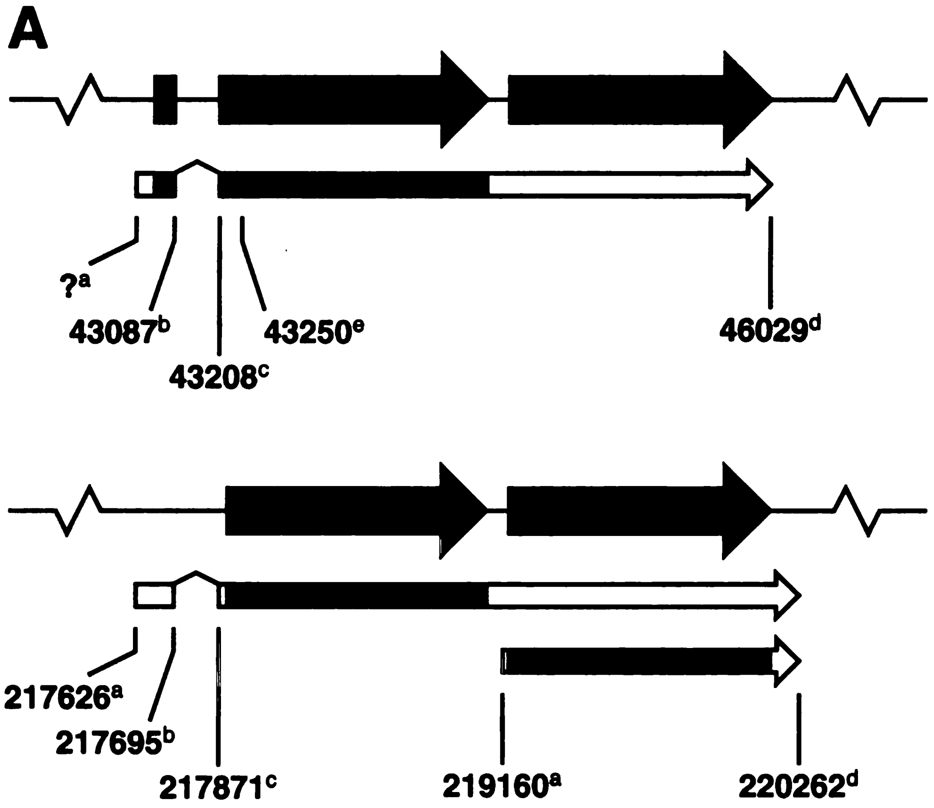
(C

(A

(T

4.1 Sequence and Transcription Analysis of CMV Chemokine Receptors

The first characterization of a virus-encoded GPCR was reported in conjunction with the cloning and characterization of the human chemokine receptor CCR1 by Neote et al. (NEOTE et al. 1993) and Gao et al. (GAO et al. 1993). They found that the amino acid (aa) sequence derived from US28 was 33% identical to that of CCR1, but also shared 32% identity to the sequences of both CXCR1 and C XCR2. When they c loned US28, they found that, due to an error in the original GenBank sequence, the predicted C-terminal of the protein was actually 65 aa, rather than 23 aa, in length, resulting in an overall length of 365aa for the US28 gene product (pUS28). This was ultimately confirmed by several other research groups (BILLSTROM et al. 1998; GAO MURPHY 1994; KUHN et al. 1995). Cloning of US27 and US28 has been done using genomic DNA from CMV. This approach does not take into consideration putative splicing of the US27- or US28-specific transcripts. In order to generate US27 and US28 expression constructs, we obtained US27-specific and US28-specific cDNAs from Toledo CMV-infected fibroblasts, rather than clones of genomic DNA. By sequencing these cDNA clones, the potential 5' and 3' ends of the US27 and US28 mRNAs were determined (Fig. 2.3A). The US28 gene does not harbor any introns. Surprisingly, the US27-specific transcript was found to be spliced within the 5' untranslated region (UTR) (Fig. 2.3A). The relevance of this splicing event has not been investigated, but it suggests that US27 expression might be regulation at the post-transcriptional level. The UL33-specific transcript was also shown to be spliced (Fig. 2.3A) (DAVIS-POYNTER et al. 1997). This splicing, however, results in a transcript containing an UL33 ORF that is different from the UL33 ORF predicted by Chee et al. (CHEE et al. 1990b). Both US27 and US28 share a common polyadenylation signal (AATAAA), of which the first two adenosines are also part of the US28 stop codon (TAA) (Fig. 2.3A). The aa coding content within the cDNA sequences of the Toledo



B

```
AD169 pUS27    M T T S T - - N N Q T L T Q V S N M T N H T L N S T...
               | | | | |                 | | | |   | | | | | | | |
Toledo pUS27   M T T S T T T T T N I M L Q V S N V T N H T L N S T...
               | |     | | |                                     |
Toledo pUS28   M T P T T T T A E L T T E F D Y D E A A T P C V F T...
```

UCSF LIBRARY

Fig

ence

the

(bla

box

and

nuc

num

code

the

Tole

Figure 2.3. Sequence analysis and transcription of the CMV UL33, US27- and US28-encoded chemokine receptors. (A) Splicing of UL33- and US27-specific transcripts. In the diagram, the chemokine receptor open reading frames (ORFs) on the CMV genome (black lines) are indicated as black arrows and the transcripts (white arrows) as black boxes. The positions of the transcription start (a), splice donor (b), splice acceptor (c), and transcription termination (d) are indicated by numbers that correspond to the nucleotide positions of the CMV AD169 genomic sequence deposited under GenBank number NC_001347. The number indicated by (e) denotes the initially predicted start codon of the UL33 ORF. (B) Alignment of the predicted aa sequences correponding to the extracellular N-termini of the chemokine receptors encoded by CMV AD169 US27, Toledo US28 and Toledo US28.

UCSF LIBRARY

strain was compared to the US27 and US28 aa sequences of the AD169 and Towne strains. The aa sequence derived from Toledo US28 differed by only two residues when compared to the US28 sequence of both AD169 and Towne. However, higher sequence variability was found for the US27-derived aa sequence. The AD169 US27-derived aa sequence differs by 14 residues from that of Towne US27 and 15 residues from that of Toledo US27. Moreover, both the potential Toledo and Towne US27-specific N-termini have two additional aa residues compared to the AD169 US27-specific N-terminus. The highest sequence variability between AD169 US27, on the one hand, and Towne and Toledo US27, on the other, was found within the potential N-terminal region of the receptor (Fig. 2.3B). An important component in binding of chemokines to their receptors is the interaction of the chemokine with the N-terminal of the chemokine receptor (reviewed in (BAGGIOLINI et al. 1997)). Whether the differences in N-terminal sequences among pUL27s of AD169, Towne, and Toledo reflect differences in chemokine binding affinity among the different CMV strains, remains to be investigated.

Transcription of each of the UL33, UL78, US27 and US28 genes is initiated at different times post-infection (pi). Transcripts of UL33, 3.3 kb in length, are detected by Northern blot analysis as early as 4 h pi and become more abundant during the late phase of infection (BODAGHI et al. 1998; DAVIS-POYNTER et al. 1997). However, inhibition of viral replication with phosphonoacetic acid (PAA) for 2 or for 7 days pi prevented detection of UL33 transcripts by Northern blot (DAVIS-POYNTER et al. 1997; WELCH et al. 1991). Interestingly, UL33-specific transcripts could be detected in infected cells treated with cycloheximide (CHX) (DAVIS-POYNTER et al. 1997). UL78-specific transcripts were detected in fibroblasts exclusively at the early stage of infection, as determined by microarray analysis (CHAMBERS et al. 1999). However, a similar gene found in the rat CMV genome, R78, was shown to be transcribed not only early, but also during the late phase of infection, as demonstrated by Northern blot analysis (BEISSER et al. 1999). The US27 gene is transcribed as a 2.9-kb mRNA only at late times (>48 hr)

UCSF LIBRARY

afe

bot

al.

foll

con

CH

mor

infa

CM

(Ko

mor

thes

fibr

US

very

(DA

oth

a be

4.2

The

bee

poly

al.

wer

wel

after infection (BODAGHI et al. 1998; WELCH et al. 1991). The US28 gene is transcribed both at early (8 h pi) and late times after infection by Northern blot analysis (BODAGHI et al. 1998) and at immediate early times (2h) pi as detected by reverse transcription (RT) followed amplification by polymerase chain reaction (PCR) (ZIPETO et al. 1999). In contrast to UL33 transcription, it was found that US28 transcription was not inhibited by CHX treatment. Furthermore, US28-specific transcripts can be found in peripheral blood mononuclear cells (PBMCs) *in vivo* (PATTERSON et al. 1998), as well as in a CMV-infected pre-monocyte cell line THP-1 *in vitro* (ZIPETO et al. 1999). Both US28- and CMV-latency-related transcripts (CLTs) from the major immediate early (MIE) locus (KONDO MOCARSKI 1995) were detected by RT-PCR in CMV Toledo-infected, THP-1 monocytic cells 7 days pi. Infectious virus could not be recovered from supernatants of these cells, but virus could be reactivated following 2 weeks of co-culture with MRC-5 fibroblasts (BEISSER et al. 2001). These findings suggest that, like MIE-derived CLTs, US28 is transcribed in latently infected cells. Since transcripts from UL33 were found at very early time points pi, similar to detection of immediate early US28 transcripts (DAVIS-POYNTER et al. 1997), it might be worthwhile determining whether UL33 and other immediate-early genes are transcribed during latency. This could eventually lead to a better understanding of gene regulation during latent CMV infection.

4.2 Expression of CMV-Encoded Chemokine Receptors

The investigation of CMV-specific chemokine receptor detection at different times pi has been frustrated by a lack of specific antibodies to these proteins, with one exception -- polyclonal antibodies were developed against a UL33 C-terminal peptide by Margulies et al. (MARGULIES et al. 1996). Using these antibodies, UL33-encoded receptors (pUL33) were detected in CMV virions, dense bodies and non-infectious enveloped particles, as well as in intracytoplasmic inclusions. The presence of pUL33 on virions and dense

UC
Franci
BRARY
UC
Franc
LIBRAR
UC
Fra
IBRA

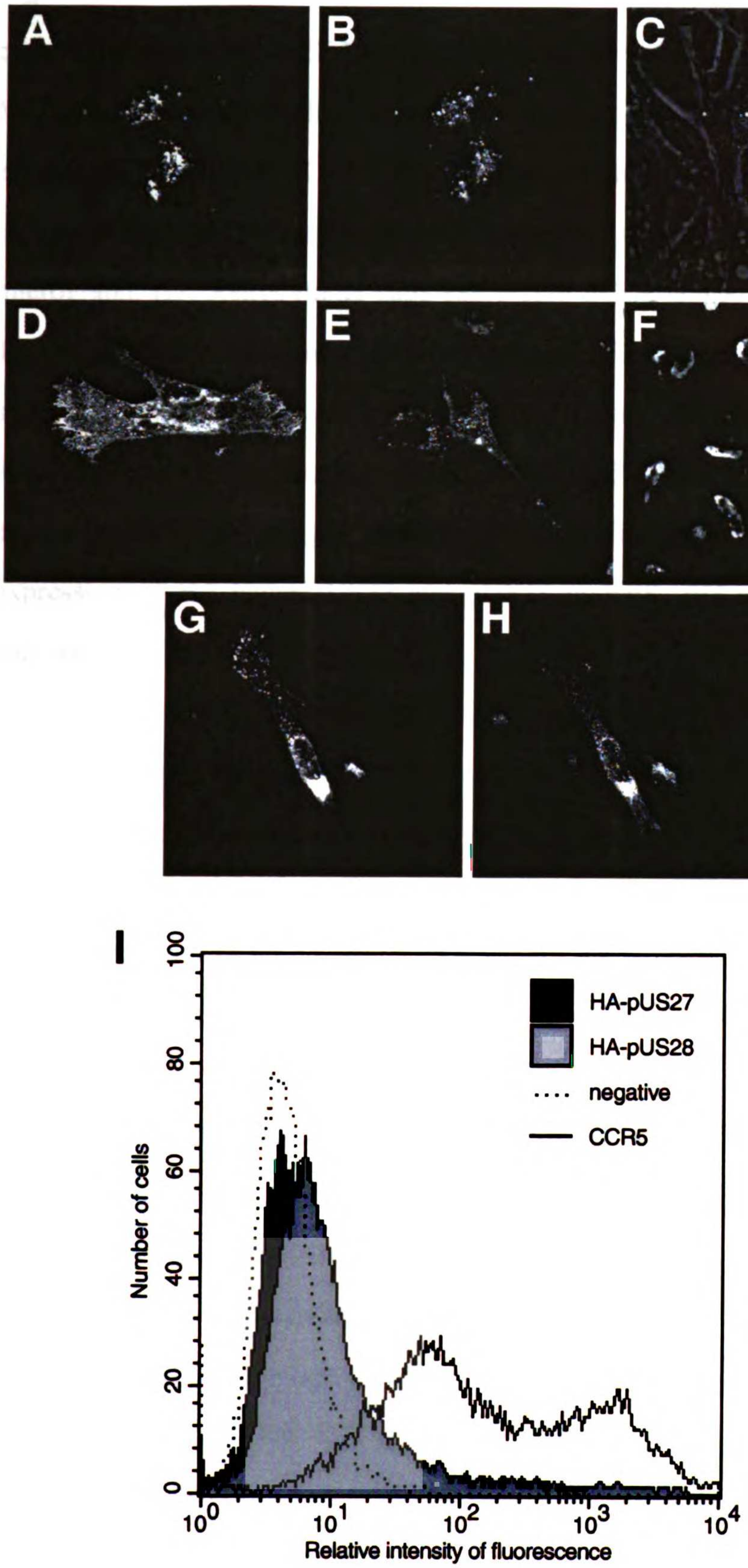
bodies led to several speculations: (i) pUL33 could participate in viral adsorption by attaching to its natural ligand(s) expressed by specific cell types, (ii) pUL33 may be disposed at the cell surface upon virus adsorption and penetration, where it could play a role in very early cell activation which would augment viral infection, and (iii) other CMV-specific chemokine receptors, if similarly incorporated into the envelopes of virions and dense bodies, could also participate in viral entry and/or host cell activation.

Expression of the putative UL78 gene product (pUL78) has not yet been reported. However, a similar receptor, encoded by the human herpesvirus 6 (HHV-6) gene U51 (pU51), was shown to accumulate in the ergastoplasm of HEK 293 and 143tk^{-} cells following transfection (Menotti et al. 1999). This localization appeared to be cell type-dependent. The pU51 receptor localizes to plasma membranes in T cells, which is a permissive cell type for HHV-6 (Menotti et al. 1999).

Determining the localization of the gene product of US27 (pUS27) and of pUS28 within infected and transfected cells has relied on the adjunction of different peptide tags such as N-terminal c-myc (Pleskoff et al. 1997), N-terminal FLAG (Streblow et al. 1999), C-terminally-tagged enhanced green fluorescent protein (EGFP) or an N-terminally-tagged hemagglutinin-specific peptide (HA) (Bodaghi and Beisser, unpublished result). Using expression vectors containing either HA- or GFP-tagged US27 or US28 genes, we were able to localize these receptors in both transiently and stably transfected cells. Cell types used include an astrocytoma cell line (U373 MG), HEK 293 and an erythrocytoma cell line (K562). The receptors have a marked tendency to be localized within the perinuclear cell center of U373 MG cells. When U373 MG cells were co-transfected with a chemokine receptor gene tagged either at the N- or the C-terminus, confocal microscopy showed that US27-EGFP and HA-US27 expression constructs resulted in co-localization of their respective gene products (Fig. 2.4A--C). Similar results were obtained for US28 expression (not shown), indicating that the presence of either a C- or N-terminal tag does not differentially affect localization of the US27- and US28-encoded receptors. When

cells transfected with either a tagged US27 or US28 expression vector were subsequently infected with Toledo CMV (Fig. 2.4D--F), several observations were made: (i) there was enhanced expression of the transfected receptor, which is not surprising in light of their being driven by the MIE CMV promoter/enhancer, (ii) there was no change in the subcellular location of tagged receptors following infection and (iii) transfection of receptors did not render astrocytoma cells resistant to infection. Similarly, upon cotransfection of astrocytoma cells with expression vectors containing either HA-US27 and US28-EGFP, or *vise versa*, US27-EGFP and HA-US28, the respective gene products colocalize (Fig. 2.4G--H). This suggests that both pUS27 and pUS28 are expressed in the same subcellular compartments in astrocytoma cells. Finally, it was reported that the pUS28 receptor could be expressed in aorta smooth muscle cells (SMC) by recombinant adenovirus containing an N-terminal FLAG-tagged US28 gene (STREBLOW et al. 1999). In these cells, the receptor adopted a polarized distribution and it is presumed that the receptor appears at the cell membrane. Recent immunofluorescent studies (FRAILE-RAMOS et al. 2001) in Hela and Cos cells demonstrated that the majority of pUS28 is within endosomes, while only 20% localizes to the cell surface.

Although many chemokine binding and signaling studies have been performed with pUS28, and transcription of the US28 gene has been confirmed in their respective expression systems (BILLSTROM et al. 1998; BODAGHI et al. 1998; GAO MURPHY 1994; NEOTE et al. 1993; VIEIRA et al. 1998), direct evidence for cell surface expression of pUS27 and pUS28 has been reported only by Pleskoff et al. (PLESKOFF et al. 1997) in transiently transfected Hela and HEK 293 cells. The cell surface expression of both pUS27 and pUS28 is significantly lower compared to that of human cellular chemokine receptors. A comparative example is shown in Fig. 2.4 I, where HEK 293 cells were transfected with vectors containing either US27 or US28, each tagged with an N-terminal, HA-encoding sequence, or with a vector containing the CCR5 receptor. Stabilization of HA-US27 and HA-US28 in U373 MG or K562 cells with a selective



UCSF LIBRARY

Figure

chemok

MG. (

express

tagged

microg

Toledo

showin

Astroc

expres

expres

analys

Figure 2.4. Subcellular localization of the CMV Toledo US27- and US28-encoded chemokine receptors (pUS27 and pUS28, respectively) in the astrocytoma cell line U373 MG. (A) An immunofluorescence micrograph (rhodamine staining) of astrocytes expressing HA-tagged pUS27. (B) The same field showing the expression of EGFP-tagged pUS27. (C) The same field combined with the corresponding bright field micrograph. (D) Astrocytoma cells expressing EGFP-tagged pUS27. (E) Human CMV Toledo-infected astrocytoma cells expressing EGFP-tagged pUS27. (F) The same field showing CMV-infected cells expressing major immediate early antigens. (G) Astrocytoma cells expressing HA-tagged pUS28. (H) The same field showing the expression of EGFP-tagged pUS27. All magnifications are × 640. (I) Cell surface expression of HA-tagged US27 and pUS28 in HEK 392 cells determined by FACS analysis.

UCSF LIBRARY

age

env

In

(P

ex

da

to

co

ex

4.

R

Bi

re

U

w

su

de

re

w

w

h

by

ch

a

co

pr

agent and subsequent cell sorting of cells expressing HA epitopes failed to result in an enrichment of HA-US27- or HA-US28-expressing cells (Beisser et al., unpublished data). In addition, HEK 293 cells expressing myc-tagged pUS28 could not be stabilized (Pleskoff *et al.*, personal communication). However, US27 and US28 could be stably expressed in U373 MG cells that stably express CMV IE1 (Beisser et al., unpublished data). This suggests that (i) both pUS27 and pUS28 inhibit cell growth and might even be toxic to the cell and (ii) that this possible growth inhibitory effect or toxicity can be compensated for by the presence of IE1 proteins. Currently, relationships between pUS28 expression and induction of cell death are under investigation.

4.3 Chemokine Binding and Signaling Properties of CMV-Encoded Chemokine Receptors

Binding of chemokines to the gene products of either UL33 or UL78 has not yet been reported. Moreover, fibroblasts infected with a CMV mutant, from which both US27 and US28 are deleted, failed to internalize RANTES or deplete extracellular MCP-1, whereas wild-type (wt) CMV was able to internalize both chemokines (Bodaghi et al. 1998). This suggests that neither UL33 nor UL78 are involved in RANTES internalization or MCP-1 depletion. In contrast, similar receptors encoded by the HHV-6 genes U12 and U51, respectively, were shown to bind several CC chemokines. Cells transfected with U12 were shown to bind RANTES, MIP-1α, MIP-1β and MCP-1 (Isegawa et al. 1998), whereas cells transfected with U51 bind RANTES, eotaxin, MCP-1, -3 and -4, as well as human herpesvirus 8 vMIP-II (Penfold et al. 1999). Additionally, the receptor encoded by U12 was shown to induce Ca^{2+} signaling upon stimulation by the aforementioned chemokines. Thus, although the genomic positions of the CMV UL33 and UL78 genes and the HHV-6 U12 and U51 genes are conserved, respectively, it is possible that the corresponding gene products of the respective betaherpesviruses have different functional properties.

UCSF LIBRARY

UC
Franc
BRARY
UC
Franc
IBRAR
UC
Fra
RA

The chemokine-binding property of pUS27 is not well characterized. However, it was shown that cells infected with a US28--deletion mutant of CMV could bind and internalize RANTES (BODAGHI et al. 1998). In contrast, RANTES binding and internalization could not be detected in cells infected with a mutant CMV strain from which both US27 and US28 were deleted. This suggests that pUS27 can bind RANTES. However, this has not yet been confirmed by conventional ligand binding studies.

The US28-encoded receptor is at present one of the most extensively studied viral chemokine receptor. It binds RANTES, MIP-1α and β, MCP-1 and 3 (BILLSTROM et al. 1998; BODAGHI et al. 1998; GAO MURPHY 1994; NEOTE et al. 1993; VIEIRA et al. 1998), but not the CXC chemokine IL-8 (BILLSTROM et al. 1998; GAO MURPHY 1994; NEOTE et al. 1993). Table 2.2 gives binding affinities of CC chemokines as determined in both US28-transfected and CMV-infected cells. It appears that, in general, RANTES and MIP-1α have higher affinities for US28 than do the chemokines MIP-1β, MCP-1 and MCP-3 (see references in Table 2.2). In addition, US28 displays high affinity for the soluble form, and possibly also for the membrane-bound form, of the CX3C chemokine fractalkine (HASKELL et al. 2000; KLEDAL et al. 1998). pUS28 expressed in Cos-7 and Hela cells is constitutively active (CASAROSA et al. 2001; FRAILE-RAMOS et al. 2001) and in Cos cells (CASAROSA et al. 2001) increases inositol-3-phosphate (IP$_3$) production by activating phospholipase C via Gαq/11. RANTES and MCP-1 stimulate IP$_3$ production further, but this activity is, however, partially inhibited by fractalkine, which therefore acts as a partial inverse agonist. Additionally, US28-transfected Cos-7 cells show constitutive activation of NFκB via Gαq/11 and Gβ/γ subunits, which is again partially inhibited by fractalkine. Neither IP$_3$ production, nor NF-κB activation could be inhibited by PTX, confirming their G$_{\alpha i}$-independent activation.

UCSF LIBRARY

Table 2.2. Chemokine Binding to CMV US28

Cell System	Ligand(s)	Kd (nM)	Reference
HEK 293 cells (transiently expressing)	MIP-1α	≈ 1[a]	(Neote et al. 1993)
K562 cells	MCP-1	6.1	(Gao & Murphy
(stably expressing)	MIP-1α	2.5	1994)
	MIP-1β	5.1	
	RANTES	3.4	
Cos 7 cells (transiently expressing)	MCP-1	0.46	(Kuhn et al. 1995)
	RANTES	0.17	
HEK 293	RANTES	~10	(Billstrom et al. 1998)
Cos 7 cells	Soluble CX3C	0.29--0.51[b]	(Kledal et al.
(transiently expressing)	Soluble CX3C with mucin	2.8	1998)
	stalk	0.748[b]	
	MCP-1	0.608[b]	
	MIP-1α	0.708[b]	
	MIP-1β	0.49[b]	
	RANTES		
CMV-infected HUVEC	RANTES	10	(Billstrom et al. 1998)
CMV –infected	MIP-1α	0.75[b]	(Bodaghi et al.
fibroblasts	MIP-1β	0.75[b]	1998)
	RANTES	0.75[b]	
	MCP-1 and 3	5x[c]	

[a] Neote et al. ((NEOTE et al. 1993)) report 2 binding affinities for MIP-1α, the second being ≈380nM

[b] These were given as IC_{50}. Conversion to Kd was done using the formula: Kd =IC_{50}. – concentration of radioactive ligand reported by the authors. Note that Kd

[c] The authors merely say that 5 times-higher concentrations of MCP-1 and 3 were required to compete the same amount of ^{125}I-MIP-1α.

UCSF LIBRARY

In human cells, some CC chemokines which can bind to US28 (MCP-1 and -3, MIP-1α) stimulate arachindonic acid (AA) release in association with phosphorylation of cytosolic phospholipase A2 (cPLA2) (LOCATI et al. 1996). Some of the very early metabolic changes in fibroblasts infected with active CMV involve stimulation of AA release (reviewed in (ALBRECHT et al. 1989)), which depends on a PTX-sensitive, phosphorylated cPLA2 chain of events (SHIBUTANI et al. 1997). This chain of events consisted of (i) phosphorylation, membrane mobilization and activation of cPLA2, (ii) concomitant increase in AA release and increase of cyclooxygenase levels, and (iii) translocation of NFκB to the nucleus (SPEIR et al. 1998; ZHU et al. 1997). It was shown earlier by Speir et al. (SPEIR et al. 1996) that CMV infection also induces reactive oxygen species (ROS), which are involved in this cPLA2 to NF-□B translocation pathway. The early induction of RANTES by CMV infection could stimulate these events in cells bearing CCRs responsive to RANTES. If pUS27 or pUS28 are structural components of the CMV envelope, similar to what has been shown for the UL33 gene product (MARGULIES et al. 1996), these receptors could be deposited by the viral envelope on the cell membrane at the time of viral entry. US28, deposited on the cell membrane by incoming viral elements or expressed at immediate early times (ZIPETO et al. 1999), might play a role in NF-κB translocation and subsequent gene activation (YUROCHKO HUANG 1999).

CMV infection of fibroblasts results in sustained activation of the MAP kinases, ERK1, ERK2 and p38, which presumably play a role in the phosphorylation of transcription factors important for CMV replication (CREB, AP-1, etc.) (BRUENING et al. 1998; RESCHKE et al. 1999; RODEMS SPECTOR 1998). In this respect, it is interesting that RANTES stimulation of US28 stably expressed in HEK 293 cells resulted in activation of ERK2, which was sensitive to inhibition with PTX (BILLSTROM et al. 1998); this activity was greater in HEK 293 cells co-transfected with $G_{\alpha16}$ protein. Activation of MAP kinases can be stimulated through chemokine receptors coupled to □ subunits of G_s, G_q

UC
San Franc
LIBRARY
UC
San Franc
LIBRAR
CALIFORNIA
UC
San Fran
LIBRA

and G_i families, as well as via $\beta\gamma$ subunits (Faure et al. 1994; Selbie Hill 1998). Although RANTES induction appears to be concomitant to MAP kinase activation, MCP-1 production is often constitutive in uninfected cell cultures (Bodaghi et al. 1998; Streblow et al. 1999) and can also activate US28. Finally, in CMV infected cells, endogenous Ca^{2+} levels increase with time after infection (Garnett 1979). One can wonder if the continuous stimulation of US28 by MCP-1 and high concentrations of RANTES, which has been shown to mobilize calcium in infected and transfected cells (Billstrom et al. 1998; Gao Murphy 1994; Neote et al. 1993; Vieira et al. 1998), might not contribute to this elevation of Ca^{2+} levels. This could be additional to Ca^{2+} signaling associated with IP_3 production mediated by the activity of pUS28 (Casarosa et al. 2001).

The consequences of chemokine-mediated activation of host cells would depend on the extent of viral replication within a given cell. Abortive infection would presumably lead to induction of CC and CXC chemokines, while full viral replication would be more likely to decrease ambient CC chemokine concentrations.

4.4 Modulation of Host Cell Chemokine Production During CMV Infection

The production of chemokines of the host organism is regulated at both transcriptional and the post-transcriptional levels. This occurs upon stimulation with cytokines in an inflammatory situation, such as during viral infection. In addition to the production of virus-encoded chemokines and chemokine receptors, it was shown that CMV infection also modulates the expression of cellular chemokines of both the CXC and CC families. Infection of human fibroblasts with any laboratory strain of CMV, as well as clinical isolates, upregulates constitutive production of IL-8 (Craigen Grundy 1996; Craigen et al. 1997; Murayama et al. 1997). We have studied IL-8 production following infection of bone marrow (BM) myofibroblasts isolated from human BM. Constitutive

IL-8 production by uninfected cells was high (ranging from 4 to 57 ng/ml) and was not modified in 12/13 BM myofibroblast cultures infected with either AD169 or Toledo strains of CMV (Michelson & Charbord, unpublished results). In contrast, AD169 strain and endothelial cell-adapted clinical isolates of CMV upregulate IL-8 production in endothelial cells (ALMEIDA-PORADA et al. 1997; GRUNDY et al. 1998). CMV infection of fibroblasts has also been shown to increase extracellular production of RANTES (MICHELSON et al. 1997), as well as MCP-1 secretion (HIRSCH SHENK 1999), at early times of infection. MIP-1α production increases in supernatants of CMV-infected, global BM stroma cultures (LAGNEAUX et al. 1996).

These modulations of chemokine production following CMV infection may be indirect, through induction of inflammatory cytokines (TNF-α, IL-1β, IFNγ, and IFNβ). Prior cytokine induction was partially controlled for in some studies. The induction of IL-8 expression in infected fibroblasts was not the result of the presence of TNF-α or IL-1 (CRAIGEN GRUNDY 1996), while stimulation of IL-8 production in endothelial cells might have been related to IL-1 and IL-6 (ALMEIDA-PORADA et al. 1997). Induction of RANTES in fibroblasts could not be attributed to the presence of the TNFα or IL-1β (MICHELSON et al. 1997). However, in subsequent studies, RANTES secretion by infected fibroblasts was reduced by 60% in the presence of IFN-β-neutralizing antibodies (Bodaghi et al., unpublished results).

In contrast to their upregulation during the early phase, at later time after CMV infection of fibroblasts and endothelial cells, CC chemokine excretion is drastically reduced. Ligand binding to chemokine receptors leads to internalization of the ligand--receptor complex, destruction of the bound ligand and subsequent recirculation of the receptor to the surface. Through this process, pUS28 has been shown to withdraw chemokines from the supernatants of infected fibroblasts (BODAGHI et al. 1998; VIEIRA et al. 1998), endothelial (BILLSTROM et al. 1998; BILLSTROM et al. 1999; RANDOLPH-HABECKER et al. 1997) and astrocytoma cells (Michelson et al., unpublished results). Infection of

fibroblasts with laboratory or clinical CMV isolates results in the disappearance of RANTES from culture supernatants starting 16hr to 24hrs after infection (MICHELSON et al. 1997). RANTES can be seen to accumulate intracellularly concomitant to its disappearance from supernatants. Exogenous, biotinylated RANTES added to infected cells 48 hr to 72 hr pi can be detected within cells after a 3 hr-adsorption when cells are infected with either the US27 or US28 null mutants of CMV (BODAGHI et al. 1998), but not when they are infected with a combined US27/US28 null mutant. The pUS28 receptor appears to have a considerable capacity for chemokine internalization, for it can simultaneously deplete RANTES and constitutively produced MCP-1 from supernatants of infected cells.

It has been reported recently that CMV infection down-regulates transcription of the gene encoding MCP-1 in fibroblasts, as detected by Northern blot analysis (HIRSCH SHENK 1999). However, in our laboratory (BODAGHI et al. 1998), infection of fibroblasts with a mutant CMV deleted of both US28 and US27 did not affect constitutive production of MCP-1 in fibroblasts, suggesting that the down-regulation of MCP-1 gene transcription is associated with either pUS27 or pUS28 expression. Somehow, simultaneous binding of RANTES and MCP-1 to pUS28 may have a feedback effect on the transcription of chemokine genes. This notion is supported by the findings of Milne et al, (MILNE et al. 2000), who studied RANTES binding to HHV-6 U51. In this system, RANTES-specific transcripts were reduced 10-fold in cells transfected with U51 expression vectors, while transcripts of β-actin and IL-8 were not affected. A similar feedback mechanism has not been described for cellular GPCRs to our knowledge.

4.5 The Implication of US28 in Retroviral Infection in Vitro

The US28-encoded chemokine receptor can serve as a co-receptor for human immunodeficiency virus (HIV) entry and play a role in cell-to-cell fusion between cells

expression HIV envelopes and those expressing pUS28 (CHOE et al. 1998; OHAGEN et al. 2000; PLESKOFF et al. 1997; RUCKER et al. 1997). The US27-encoded receptor promotes neither cell fusion, nor HIV infection. Hela, U373 MG, and neuroblastoma (U87) cells co-expressing pUS28 and CD4 can be infected by some monocyte-tropic and dual-tropic HIV strains, but not by T lymphocyte-tropic HIV strains. Fusion of cells expressing pUS28 with cells expressing monocyte-tropic and, much less efficiently T-cell tropic, HIV envelopes also occurs. Thus, pUS28 behaves much like the CC chemokine receptors CCR3 and CCR5 as concerns co-receptor activity for HIV, but is much less efficient.

Co-expression of US28 with retroviral proteins other than those of HIV, such as Human T cell Lymphoma-Leukemia virus-1 gp46 and gp21, as well as Vesicular Stomatosis virus (VSV)-G proteins, also leads to increased cell-to-cell fusion (PLESKOFF et al. 1998). Various mutations within the US28 gene affect fusion with cells expressing HIV envelopes or VSV-G. Deletion of N-terminal aa 2--22 abolishes fusion with HIV envelope expressing cell, but leads to increased fusion with cells co-expressing VSV-G. Removal of the C-terminus (aa residues 296--355) has no effect on HIV co-induced fusion, but again increases fusion mediated by VSV-G. On the contrary, a point mutation in the second extracellular domain of US28 decreases its capacity to mediate cell-to-cell fusion. US28-mediated fusion was seen with human, macaque and feline cells, but not with murine or rat cells. Thus, US28 expression may contribute to transfer of CMV, HIV, and perhaps other viruses from cell to cell via fusion.

4.6 Adaptive Evolution of Human CMV Chemokines and Chemokine Receptors

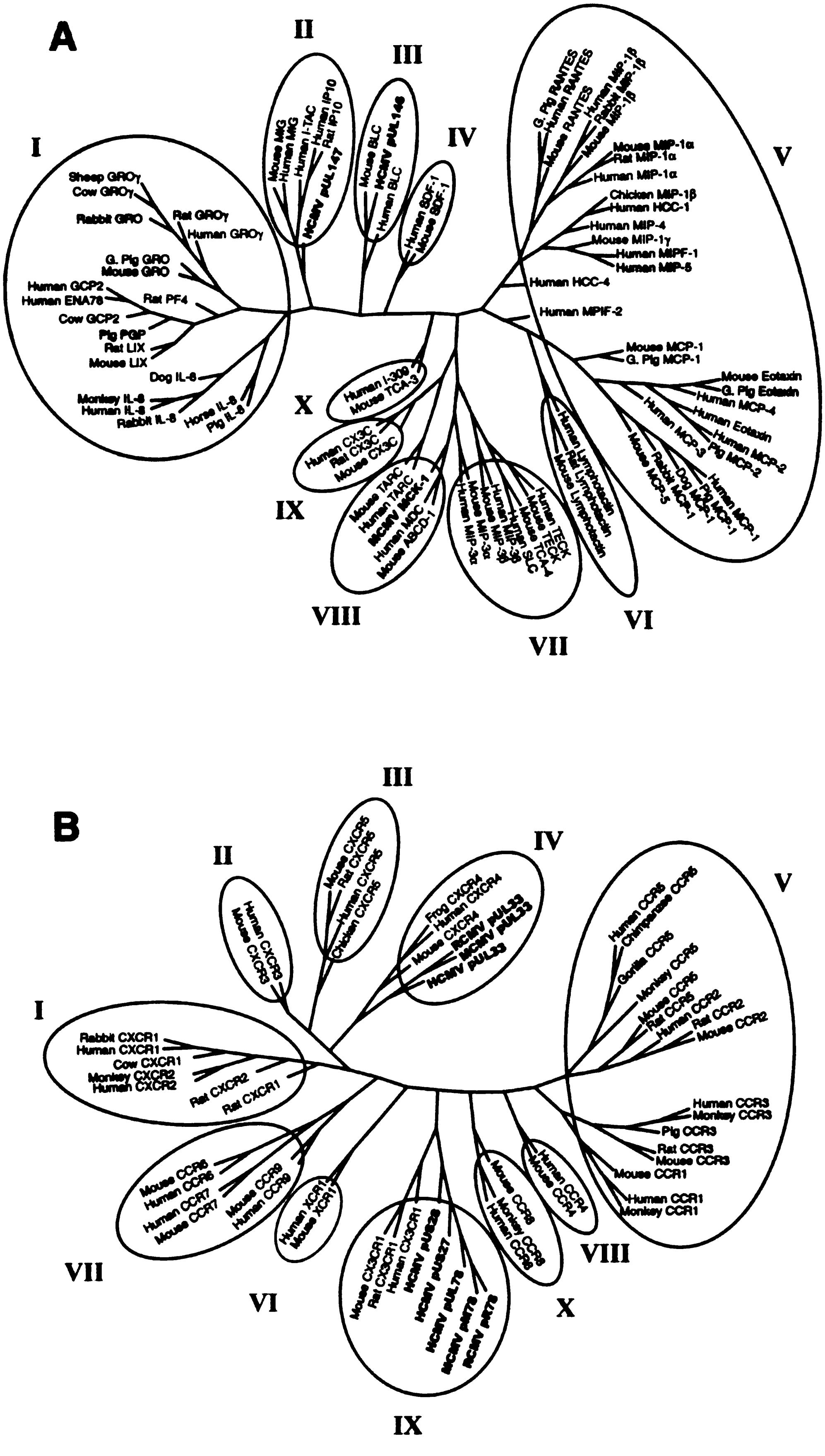
Human chemokines and their receptors have co-evolved in a correlated manner, as evidenced by the correlated patterns of clustering between evolutionary trees of well-characterized chemokine and chemokine receptors (GOH et al. 2000). Consequently, through computation, we can augment our experimental understanding of cytokine ligand--receptor preferences. By analyzing the potential co-evolution of chemokines and chemokine receptors of both human and viral origin, inferences can be made about the human protein-binding partners of the orphan CMV chemokines and receptors. Here, phylogenetic trees were constructed from the multiple sequence alignment of both chemokines (Fig. 5A) and their receptors (Fig. 5B), according to a method described by Goh et al. (GOH et al. 2000), in order to predict the probable interaction of CMV-encoded chemokines and chemokine receptors with chemokines and chemokine receptors of the host. For this purpose, both CXC chemokines encoded by HCMV UL146 and UL147, as well as a murine CMV (MCMV) ORF m131-encoding CC chemokine, MCK-1 (SAEDERUP et al. 1999), were included in the chemokine tree to determine their binding specificities. In addition, human CMV chemokine receptor sequences derived from HCMV UL78, UL33, US27 and US28, as well as both R33 and R78 from rat CMV (RCMV) and both M33 and M78 from MCMV were added to the chemokine receptor trees.

In the chemokine tree (Fig. 2.5A), MCMV MCK-1 clusters next to the MDC group and to the MIP-3α, MIP-3β, SLC, and TECK groups. This implies that MCK-1 is a CC chemokine-like protein that can potentially bind to CCR4, CCR6, CCR7, and/or CCR9. These receptors are predominantly expressed on macrophages and dendritic cells. The chemokines encoded by HCMV UL146 and UL147 cluster with CXC type chemokines. The UL146 gene product (also known as vCXC-1 (PENFOLD et al. 1999)) clusters with ligands of the CXCR5 receptor. Although the vCXC-1 chemokine was found only to bind

to CXCR2 out of an array of CCR1--CCR8, CXCR1--CXCR4, CX3CR1, and the US28-encoded receptor (PENFOLD et al. 1999), it could also be a potential binding partner for CXCR5. The UL147-encoded chemokine groups together with MIG, IP10, and I-TAC -- all ligands for the CXCR3 receptor. Therefore, we can predict that the gene product of UL147 will bind to CXCR3 or a closely related receptor.

In the chemokine receptor tree (Fig. 2.5B), CMV US28, US27, and UL78 cluster together very closely within the CX3CR group. Among the chemokine receptors, human CMV US28 has the highest similarity with CX3CR1 -- the receptor for CX3C, or fractalkine. This corresponds to known experimental findings that human CMV US28 binds CX3C (KLEDAL et al. 1998). Although US27 and UL78 are in the same cluster as US28, they appear to branch away from the rest of the chemokine receptor tree. It is possible that they can bind other CC chemokines, but it would be difficult to assign binding partners to these proteins. Finally, human CMV UL33, another orphan viral chemokine receptor, clusters quite closely to CXCR4. This suggests that CMV UL33 could bind SDF-1 or a CXC chemokine that is closely related. Taken together, these inferences on ligand-receptor specificity can possibly aid in the characterization of binding preferences of the CMV proteins.

UC
Franc
LIBRAR
UC
Franc
LIBRAR
UC
an Fra
LIBRA


A
I
II
III
IV
V
VI
VII
VIII
IX
X
Sheep GROγ
Cow GROγ
Rabbit GRO
Rat GROγ
Human GROγ
G. Pig GRO
Mouse GRO
Human GCP2
Human ENA78
Rat PF4
Cow GCP2
Pig PGP
Rat LIX
Mouse LIX
Dog IL-8
Monkey IL-8
Human IL-8
Rabbit IL-8
Horse IL-8
Pig IL-8
Mouse MIG
Human MIG
Human I-TAC
Human IP10
Rat IP10
HCMV pUL147
Mouse BLC
HCMV pUL146
Human BLC
Human SDF-1
Mouse SDF-1
G. Pig RANTES
Human RANTES
Mouse RANTES
Human MIP-1β
Rabbit MIP-1β
Mouse MIP-1β
Mouse MIP-1α
Rat MIP-1α
Human MIP-1α
Chicken MIP-1β
Human HCC-1
Human MIP-4
Mouse MIP-1γ
Human MIPF-1
Human MIP-5
Human HCC-4
Human MIPIF-2
Mouse MCP-1
G. Pig MCP-1
Mouse Eotaxin
G. Pig Eotaxin
Human MCP-4
Human Eotaxin
Human MCP-2
Pig MCP-2
Human MCP-3
Mouse MCP-5
Rabbit MCP-1
Dog MCP-1
Pig MCP-1
Human MCP-1
Human Lymphotactin
Rat Lymphotactin
Mouse Lymphotactin
Human TECK
Mouse TECK
Mouse TCA-4
Mouse SLC
Human SLC
Human MIP-3β
Mouse MIP-3β
Mouse MIP-3α
Human MIP-3α
Mouse TARC
Human TARC
MCMV MCK-1
Human MDC
Mouse ABCD-1
Human CX3C
Rat CX3C
Mouse CX3C
Human I-309
Mouse TCA-3
B
I
II
III
IV
V
VI
VII
VIII
IX
X
Rabbit CXCR1
Human CXCR1
Cow CXCR1
Monkey CXCR2
Human CXCR2
Rat CXCR2
Rat CXCR1
Human CXCR3
Mouse CXCR3
Mouse CXCR5
Rat CXCR5
Human CXCR5
Chicken CXCR5
Frog CXCR4
Human CXCR4
Mouse CXCR4
RCMV pUL33
MCMV pUL33
HCMV pUL33
Human CCR5
Chimpanzee CCR5
Gorilla CCR5
Monkey CCR5
Mouse CCR5
Rat CCR5
Human CCR2
Rat CCR2
Mouse CCR2
Human CCR3
Monkey CCR3
Pig CCR3
Rat CCR3
Mouse CCR3
Mouse CCR1
Human CCR1
Monkey CCR1
Human CCR4
Mouse CCR4
Mouse CCR8
Monkey CCR8
Human CCR8
Mouse CX3CR1
Rat CX3CR1
Human CX3CR1
HCMV pUS28
HCMV pUS27
HCMV pUL78
MCMV pM78
RCMV pR78
Human XCR1
Mouse XCR1
Mouse CCR6
Human CCR6
Human CCR7
Mouse CCR7
Mouse CCR9
Human CCR9

Fi

By

ar

cc

ar

be

w

g

F

r

Figure 2.5. Phylogenetic trees of chemokines (A) and chemokine receptors (B). By employing a linear regression analysis on the evolutionary pairwise distances among all the proteins in the multiple sequence alignment, a correlation coefficient was calculated based on the known binding partners in the chemokine and the chemokine receptor trees. Due to the similarity of the clustering patterns between the trees, a correlation coefficient of 0.57 with a *p*-value less than 10^{-4} was obtained for the non-CMV chemokines and their receptors. The encircling of groups was is based on the branching of the chemokine receptor tree. The Roman numbers that indicate each chemokine group refer to the corresponding receptor group each of which has been numbered accordingly.

5 Putative CMV-Encoded Chemokine and Chemokine Receptor Functions

The CMV-encoded chemokine and chemokine receptors could have diverse and combined functions. These include activation of the host cell (discussed above), with or with subsequent stimulation of viral replication, viral dissemination by chemokine-regulated trafficking of infected cells, and modulation of the behavior and trafficking of cells involved in hematopoiesis and immune responses. These functions may have effects both at cellular and systemic level.

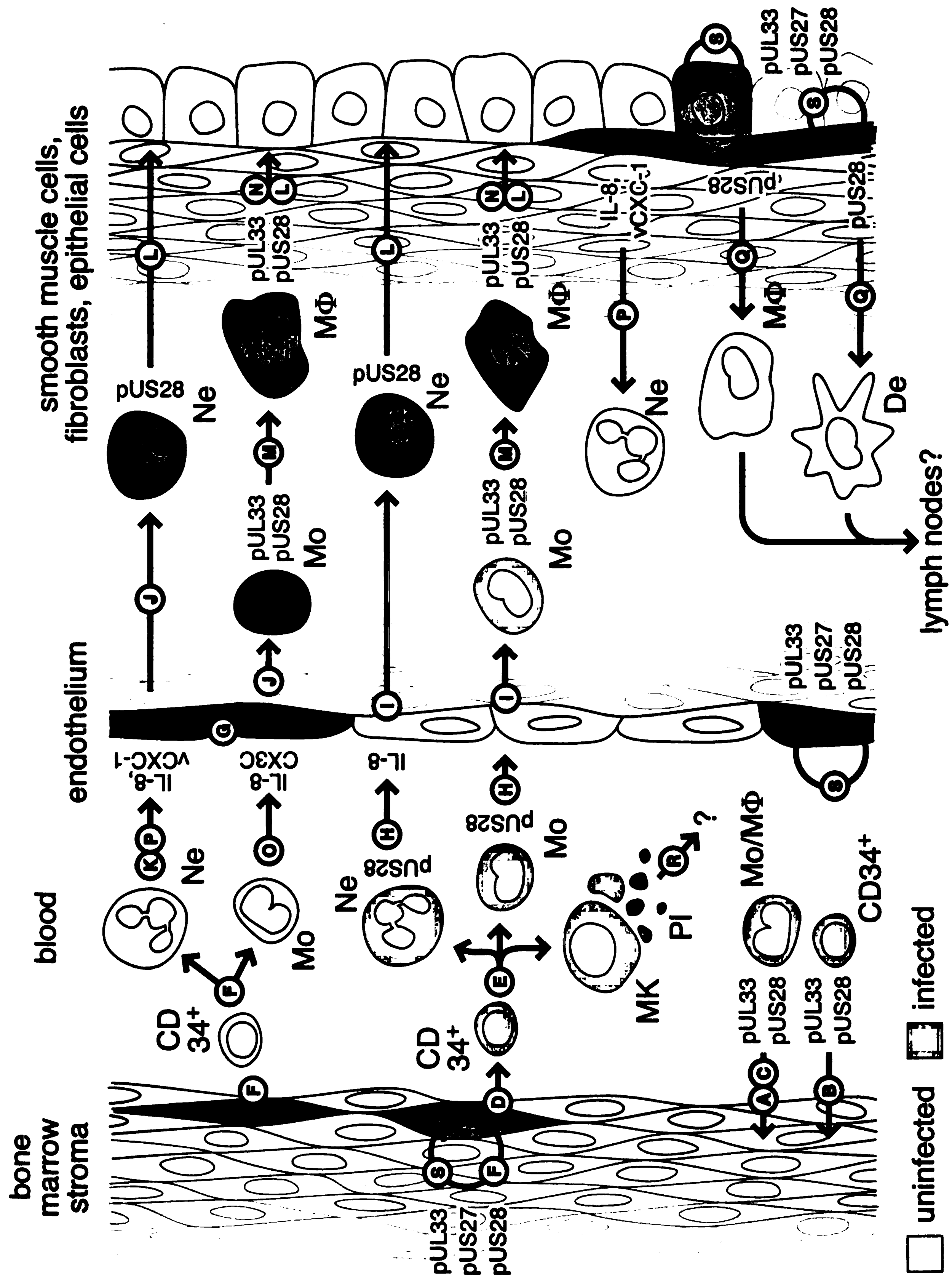
5.1 The Role of CMV-Specific Chemokine- and Chemokine Receptor in Viral Dissemination and Persistence (Figure 2.6)

Cell types that are fully permissive for CMV infection, i.e. allow full viral replication leading to excretion of new infectious particles and cell lysis, include fibroblasts, smooth muscle cells, endothelial cells, epithelial cells of the retina and excretory organs, such as salivary glands. Infection of most of these cell types is associated with immunosuppression and CMV disease. However, infection of epithelial cells from excretory organs is probably essential for virus transmission between healthy individuals. In contrast to its ability to replicate in the afore-mentioned cell types, CMV remains latent, i.e. in a non-replicative state, in myeloid cells, such as granulocyte/monocyte progenitors and mature monocytes. Possible mechanisms of trafficking of CMV in vivo between fully permissive cells, on the one hand, and cells that are latently infected with CMV, on the other, is not well understood. Considering that trafficking of myeloid cells toward inflammatory sites is mediated by chemokine receptors and the possibility of many myeloid cells being latently infected by CMV, we propose that myeloid cells shuttle CMV to fully permissive target cells. In addition, we propose that myeloid cells can take up CMV from fully permissive, infected cells. The resulting two-way traffic

may be orchestrated by virus-encoded chemokines and chemokine receptors. The putative roles of these chemokines and chemokine receptors in vivo CMV trafficking are illustrated in Fig. 2.6.

Many of the viral dissemination pathways suggested below require that transmission of virus from the infected cell to an adjacent target cell be established either via CMV-induced cell-to-cell contacts, or by shedding of virus from the carrier cell and subsequent uptake of the virus by the target cell. Some CMV-encoded proteins, predominantly structural glycoproteins such as gB and gH, were shown to play an important role in the cell-to-cell spread of CMV during infection in vitro (Baldwin et al. 2000; Bold et al. 1996; Navarro et al. 1993; Reschke et al. 1999). In addition, we have discussed the potential of pUS28 to provoke cell-to-cell fusion in association with retroviral proteins (Pleskoff et al. 1998; Pleskoff et al. 1997). Consequently, not only trafficking of CMV excreting cells, but also subsequent transmission of the virus through cell-to-cell contacts, may be mediated by pUS28.

CMV infection induces the production of cellular chemokines. In particular, IL-8 production is increased upon infection of fibroblasts (Craigen Grundy 1996; Craigen et al. 1997), endothelial cells (Almeida-Porada et al. 1997; Grundy et al. 1998) and monocytic THP-1 cells (Murayama et al. 1997). Recently, Grundy et al. [Grundy, 1998 #41] illustrated how a CMV-induced increase of IL-8 production, and perhaps Groα, could assist viral dissemination. Supernatants from CMV-infected endothelial cells contained elevated levels of IL-8 and Groα, relative to supernatants of uninfected cells; these supernatants promoted neutrophil migration across an endothelial cell barrier.



UCSF LIBRARY

Figure 2.6. Proposed mechanisms of chemokine- and chemokine receptor-dependent trafficking and persistence of CMV. The chemokines or chemokine receptors suggested to play a role in CMV trafficking are indicated in the figure adjacent to each of the cells that express these molecules. CMV-infected monocytes (Mo) expressing either pUL33 or pUS28 could infect bone marrow (BM) stromal cells expressing SDF-1 or RANTES and MCP-1, respectively (A). In the case of BM transplantation, CD34$^+$ cells, known to be latently infected in healthy donors (Kondo Mocarski 1995), might be attracted partially through a pUL33/SDF-1 or a pUS28/MCP-1 interaction (B). In the BM, alloreactivty (Soderberg-Naucler et al. 1997) following transplantation could result in the differentiation of transplanted Mo into MΦ (Sinzger et al. 1997), thereby resulting in full CMV replication in these cells with subsequent infection of stromal cells (C). Infected stromal cells (Lagneaux et al. 1995) could transmit infection to BM progenitors and assist in the establishment of latency by upregulation of chemokines which inhibit CD34$^+$ proliferation (MIP-1α) (Broxmeyer Kim 1999; Lagneaux et al. 1996), or by down-regulation of necessary stimulatory factors like SCF (Lagneaux et al. 1996) (D). Latently infected progenitors would carry the CMV genome during their maturation and liberation into the circulation (Hahn et al. 1998) (E). Mobilization of matured myeloid cells might be enhanced by pUS27/28 withdrawal of hematopoietic inhibitory factors (MCP-1 (Cashman et al. 1990), MIP-1α (Broxmeyer Kim 1999)) (F) and by increased production of IL-8 by infected endothelial cells (Craigen et al. 1997) (G). The possible expression of pUS28 on latently infected myeloid cells (Mo, neutrophils (Ne)) in the blood stream could play a role in their chemoattraction to endothelial cells expressing CX3C (H), thereby allowing both transmission of infection to endothelium and transmigration of infected cells into tissues (I). CMV transmitted to endothelial cells would become a source for new infection of transmigrating Mo and Ne (Grundy et al. 1998; Revello et al. 1998) (J). Adhesion of uninfected cells might be enhanced by expression of the CMV CXC chemokine, vCXC-1, and/or IL-8 and Groα on infected

endothelium and their interaction with CXCR2 on Ne (K). Transmigrated Ne and Mo might transmit virus to tissue epithelium, smooth muscle cells, and fibroblasts (Sinzger Jahn 1996), again via pUS28-facilitated cell fusion (L). Differentiation of latently infected Mo into tissue MΦ at sites of inflammation (M) could transmit virus to neighboring tissue components by direct infection with cell-free virus (Sinzger et al. 1996) (N). In the early stage of infection, epithelial, endothelial and smooth muscles cells could attract Mo due to CMV induction of RANTES acting on cellular receptors such as CCR1 and CCR5 (O), and later through interaction of vCXC-1 with CXCR2 on Ne (Penfold et al. 1999) (P). CMV could be tranferred from either infected smooth muscle cells, fibroblasts, or epithelial cells upon interaction of US28 with cell surface-expressed CX3C from surveilling MΦ or dendritic (De) cells (Q). Subsequently, CMV could be transported to the lymph nodes for further dissemination. Although the role for lymph nodes in CMV dissemination is unclear, CMV has been localized in these tissues (Boriskin et al. 1999). Similarly, megakaryocytes (MK) and blood platelets (Pl) could disseminate CMV (R), since it was shown that MK are susceptible to CMV infection (Crapnell et al. 2000). Finally, in addition to their function in mediating CMV trafficking, pUL33, pUS27 and pUS28 could establish persistent CMV infection in either BM stroma, smooth muscle cells, endothelium (Billstrom et al. 1998), fibroblasts (Bodaghi et al. 1998), or epithelial cells (Beisser et al. 1998). This could be established by autocrine stimulation, or constitutive activity of these receptors (Crapnell et al. 2000) . Alternatively, these receptors could act as a chemokine sink, sequestering all extracellular inflammatory chemokines in order to evade immune surveillance. Both signalling and sequestration might render the local environment favorable for CMV persistence (S).

Neutrophils that are either cocultivated with, or migrated across, infected endothelial cells take up viral products, in particular pp65 (GRUNDY et al. 1998; REVELLO et al. 1998). CMV could be reactivated by subsequent co-culture of pp65$^+$ neutrophils with fibroblasts. These observations were confirmed by Gerna et al. (GERNA et al. 2000), who showed cell fusion between neutrophils and infected endothelial cells by electron microscopy. They also reported that CMV replicated abortively in neutrophils (GERNA et al. 2000). Thus, it is likely that CMV is shuttled between fully permissive cells by neutrophils. Recently, a CMV-encoded chemokine (vCXC-1) was identified. This chemokine was shown to be a potent chemoattractant of neutrophils (PENFOLD et al. 1999). Therefore, neutrophil-mediated shuttling of CMV might be initiated by the attraction of neutrophils to infected cells expressing vCXC-1, as well as by upregulation of IL-8 and GROα, (Fig 2.6).

Many CXC chemokines that can bind specifically to CXCR2 function as inhibitors of myelopoiesis (reviewed in (BROXMEYER KIM 1999)). The CMV chemokine vCXC-1 desensitizes the cellular receptor CXCR2 expressed at the surface of neutrophils to further stimulation by NAP-2, GROα, -β or -γ, ENA-78, or GCP-2 (PENFOLD et al. 1999). The majority of these chemokines (NAP-2, GROβ, ENA-78, and GCP-2) are inhibitory to hematopoiesis. Thus, vCXC-1 can potentially interact with chemokine receptor(s) involved in myelopoiesis, although it is not yet known whether this would be stimulatory or inhibitory. It is also not known whether vCXC-1 is expressed by CMV-infected hematopoietic progenitors. If so, it could serve an autoregulatory function in which vCXC-1 would stimulate the release of CMV-harboring, differentiated myeloid cells into the circulation for further dissemination. Alternatively, it could autosuppress the differentiation of CMV-infected progenitors in the absence of other inhibitory chemokines in order to preserve latency. The putative stimulatory/suppressive effect of vCXC-1 on myelopoiesis is indicated in Fig. 2.6.

UC
Franc
LIBRAR

UC
Franc
LIBRAR
CALIFORN

LIBRAR
Franc
UC

UC
Fra
LIBRA

Previously, cells expressing pUS28 were shown to bind the CX3C chemokine, fractalkine (FK) (KLEDAL et al. 1998), interacting with many of the same epitopes of FK as does CX3CR1 (MIZOUE et al. 2001). FK exists in a soluble and a membrane-bound version. In its membrane-bound form it consists of a chemokine-like domain, a mucin stalk, a transmembrane domain and a cytoplasmic tail. Kledal et al. (KLEDAL et al. 1998) proposed a role for pUS28 in the adhesion of leukocytes latently infected by CMV to the surface of CX3C-expressing endothelial cells. Recent studies by Haskell et al. (HASKELL et al. 2000) supported this proposal. They constructed chimeras of RANTES, MIP-1α, MCP-1 and IL-8 bound to the FK mucin stalk and anchored these chimeric proteins, as well as native FK, to glass slides. Using these immobilized chimeras and FK, they showed that 300-19 cells transfected with US28 can adhere to antibody-tethered FK and become immobilized under shear flow conditions. Although cells adhered to CC chemokine chimeras under static conditions, they were not immobilized under flow-shear conditions. These results demonstrate that membrane-bound FK is potentially sufficient to immobilize CMV-infected cells in the absence of other adhesion molecules. The US28 gene is transcribed in infected peripheral blood leukocytes from CMV seropositive individuals *in vivo* (PATTERSON et al. 1998), and in a monocytic cell line, THP-1, *in vitro* (ZIPETO et al. 1999). These observations indicate that CMV-infected monocytes and possibly also monocytic progenitors may express pUS28. This implies a mechanism for CMV to traffic from monocytes to or through the endothelium either by a dhesion and subsequent cell-to-cell transmission of CMV, or by transendothelial migration of the monocytic cells into underlying tissues (Fig 2.6). Smooth muscle cells infected with CMV or transfected with US28 display chemokinesis in the presence of MCP-1 and chemotactic properties in a RANTES gradient (STREBLOW et al. 1999). Although this may reflect pUS28-mediated smooth muscle cell migration in CMV-related vascular disease *in vivo*, it is less clear what the role of migrating smooth muscle cells may have in the dissemination of CMV in healthy individuals.

In a ddition to the proposed role of pUS28-CX3C interaction in trafficking CMV from monocytes to and across endothelial, there exists another possible mode of CMV exchange between cells. Previously, it was shown that macrophages and dendritic cells express CX3C chemokines on their cell surface (Bazan et al. 1997; Imai et al. 1997). Since these cell types are involved in immune surveillance, they may encounter CMV-infected cells expressing pUS28. Cells that are fully permissive for CMV infection are likely to express pUS28 following infection *in vivo*, since the US28 gene was shown to be expressed in fibroblasts, smooth muscle cells and endothelial cells *in vitro* (Billstrom et al. 1998; Bodaghi et al. 1999; Streblow et al. 1999; Vieira et al. 1998). Hence, adhesion between infected cells and antigen-presenting cells (macrophages or dendritic cells) could result in subsequent cell-to-cell transmission from the former to the latter two cell types (Fig. 2.6)

The UL33 gene product may also play an important role in CMV dissemination. UL33 is transcribed at very early times pi (Davis-Poynter et al. 1995). Consequently, UL33 may be transcribed in latently infected myeloid cells, as are immediate early genes and US28. In addition, UL33-like genes were identified and characterized in the genomes of murine (M33) and rat (R33) cytomegalovirus (Beisser et al. 1998; Davis-Poynter et al. 1997). Mutant viruses, from which these UL33 gene homologs were deleted showed no difference in replication efficiency *in vitro*, compared to wild-type viruses. However, *in vivo*, these mutant viruses were unable either to enter or to replicate in the salivary gland epithelium of infected mice and rats. Similarly, the UL33 gene, of which both sequence and genome location correspond to those of M33 and R33, may therefore be essential for salivary gland tropism in humans. In Fig. 2.6, we propose a role for the UL33 gene product as a chemotaxis-driving factor in infected monocytes or macrophages. Similar to what was proposed for pUS28, pUL33 possibly mediates CMV trafficking by attracting latently infected cells into solid tissue, in particular the salivary gland epithelium and possibly other secretory tissues. SDF-1, a CXC chemokine that is constitutively

expressed by epithelial cells (Pablos et al. 1999) is a candidate ligand for UL33, as suggested in section 4.5. Consequently, chemotaxis of infected monocytes toward the epithelial layer could be driven by interaction of pUL33 with SDF-1 (Fig. 2.6). Alternately, the possibility remains that, in order to maintain persistent CMV infection in salivary gland epithelial cells, pUL33 may have to be expressed at the surface of these cells. Thus, an intracellular activation state could be established by continuous signaling of pUL33 by SDF-1, to establish an environment suitable for CMV persistence. This continuous signaling could occur either through uninterrupted binding of ambient chemokine, or through constitutive activity of the receptor (Fig. 2.6).

5.2 Modulation of Host Cell Chemokine Production in Relation to CMV Dissemination and Persistence

CMV infection rarely causes overt disease in immunocompetent individuals. Even in immunocompromised patients, active viral replication does not necessarily result in end-organ disease. Factors that tilt the balance between active virus replication and CMV disease are not known. It is most likely that CMV utilizes the chemokine network to propel infected cells into an environment conducive either for replication, persistence or latency. Once there, viral modulation of chemokines could assist in avoiding immune detection of the infected cell at that site.

RANTES can induce the release of IFN-γ (Appay et al. 2000), which is not only an inhibitor of many chemokines (Baggiolini 1998), but also blocks CMV replication after expression of IE proteins ((Bodaghi et al. 1999) and references therein). CMV induction of RANTES could thereby indirectly result in a persistent infection.

In the early stages of viral replication, CMV induces production of RANTES. Binding of chemokines to extracellular proteoglycans concentrates them and enhances their activity (Diasbaruffi et al. 1998; Luster et al. 1995; Oravecz et al. 1997). However,

Schaarschmidt *et al.* (SCHAARSCHMIDT et al. 1999) reported that CMV infection down-regulates proteoglycan transcription. Thus, secreted RANTES would be more likely to form a gradient around uninfected, proteoglycan-producing cells, thereby leaving infected cells "sheltered" from attack. *In vivo*, RANTES production was significantly higher in bronchoalveolar lavage (BAL) fluids during CMV pneumonitis than in lung transplant patients with non-CMV-related allograft rejection or in transplant patients without complications (MONTI et al. 1996). BAL macrophages isolated from patients with CMV pneumonitis spontaneously released more RANTES than those from control patients. This enhanced production returned to baseline with the resolution of infection. Such high production of RANTES could lead to blocking of lymphocyte cytotoxic activity (APPAY et al. 1999).

At a later stage of CMV infection, local inflammatory reactions could be controlled by chemokine down-regulation around the infected cells. The pUS28 receptor adsorbs RANTES from the infected cell environment (BILLSTROM et al. 1999; BODAGHI et al. 1998). RANTES, as well as MIP-1α/β and MCP-1 and 3, which also bind pUS28, are chemoattractant for T, dendritic and NK cells (reviewed in (LOETSCHER et al. 2000)). RANTES adsorption by pUS28 could inhibit establishment of a chemokine gradient and thereby block both lymphocyte attraction and effector mechanism activation (HADIDA et al. 1998).

The majority of CC chemokines inhibit proliferation of hematopoietic progenitors activated by cytokines (reviewed in (BROXMEYER KIM 1999)). These include MIP-1α, which is induced by infection of BM stroma (LAGNEAUX et al. 1996). Paradoxically, CMV infection would seem to down-regulate some of the inhibitory chemokines. Secretion by BM myofibroblasts of constitutively produced MCP-1, an even more potent inhibitor of progenitor proliferation (CASHMAN et al. 1990), is abolished in CMV-infected stromal myofibroblasts (Michelson & Charbord, unpublished results). This was not seen with ΔUS28 or ΔUS28/27 CMV mutants. Interaction of US28 in progenitors

with inhibitory CC chemokines could also play a role in maintaining latency/persistence by inhibiting proliferation of these cells.

In vivo, CMV DNA can be found in circulating CD34$^+$ and in BM aspirates of healthy CMV carriers (HAHN et al. 1998; HAHN MOCARSKI 1996; KONDO et al. 1994; KONDO MOCARSKI 1995; MENDELSON et al. 1996; MINTON et al. 1994; VONLAER et al. 1995). It was also detected in pretransplant trephine BM biopsies of healthy BM donors and recipients by in situ hybridization and/or immunochemical detection of CMV immediate early antigens (FEST et al. 1994a; FEST et al. 1994b; PENCHANSKY KRAUSE 1979). Viral DNA persists within progenitors throughout their differentiation and maturation (HAHN et al. 1998), particularly within the myeloid lineage (ZHURAVSKAYA et al. 1996). *In vitro* CMV infection of BM and cord blood progenitors in the absence of stromal cells causes inhibition of colony formation (reviewed in (MICHELSON 1997) and see (SINDRE et al. 2000)). Moreover, CMV has been implicated in pancytopenia following bone marrow (BM) transplantation (reviewed in (ALMEIDA-PORADA ASCENSAO 1996)). Related to this is the fact that CMV induces IL-8 production (ALMEIDA-PORADA et al. 1997; CRAIGEN GRUNDY 1996; CRAIGEN et al. 1997; MURAYAMA et al. 1997). This chemokine is a renown mobilizer of CD34+ progenitors into the circulation and could thus play a role in depletion of progenitors from BM. Increased serum levels of IL-8 were found to correlate with CMV infection and antigenemia after BM transplantation (FIETZE et al. 1994; HUMAR et al. 1999). IL-8 plasma levels were also significantly increased, while MIP-1α levels decreased, in renal transplant patients who later developed CMV disease (NORDOY et al. 1999). Here, CMV-mediated mobilization of progenitor cells by IL-8 up-regulation could play a significant role in the dissemination of latently infected progenitors.

6 Conclusions

From what is known about CMV-encoded chemokines and chemokine receptors, it appears that their participation in immune evasion would be mainly at the level of viral

dissemination sheltered from the immune system through (cell-to-cell) passage and movement of receptor bearing infected cells bi-directionally across endothelial barriers. In addition, the ability of pUS28 to withdraw CC chemokines from the environment of infected cells could also confer a measure of immune evasion by blunting effector lymphocyte migration and activation.

So far, there have been reports for up and down-regulation of chemokines and cytokines of the host organism at least at the transcriptional level, and by chemokine scavenging via CMV-encoded chemokine receptors. In addition, CMV may contribute to the effects of chemokine/cytokine modulation by expressing viral chemokines. Each of the CMV-encoded chemokine and chemokine receptor genes may exert individual functions in either dissemination or the establishment and maintenance of viral latency *in vivo*. Several of these putative functions are outlined in this chapter. However, there may also be an intricate interplay between the different cytokines, chemokines and chemokine receptors of both viral and host origin (SELBIE HILL 1998). For this, we still need to examine especially the kinetics of expression of the CMV-encoded pUL33, pUL78, pUS27, pUS28, vCXC-1 and the putative chemokine encoded by ORF UL147 in more diverse cellular environments than those that have been studied to date. Special attention should be paid to cytokine/chemokine interactions in CMV-infected cells of the myeloid lineage. Although these cells are in general not permissive for full CMV replication, they are important CMV carriers that are most likely steered by the complex cytokine/chemokine network and probably play an important role in viral dissemination within and between individuals.

UC
Franci
LIBRAR

C

ot

Bo

da

w

w

1

p

e

Cell-to-cell fusion coupled with CMV-induced down-regulation of HLA molecules (see other chapters in this book) and withdrawal of chemokines (Billstrom et al. 1999; Bodaghi et al. 1998; Vieira et al. 1998) would allow infected cells to avoid immune detection. Full, active viral replication *in vivo* seems to occur at limited, confined sites within target organs. Effectively, the CMV genome can be detected in many organs and within many cell types (Hendrix et al. 1997; Myerson et al. 1984; Toorkey Carrigan 1989), but expression of late antigens (Sinzger et al. 1997) with the development of pathology is rare compared to the incidence of genome-carrying cells detected (Hendrix et al. 1997; Larsson et al. 1998).

References:

Albrecht T, Boldogh I, Fons M, Lee C H, AbuBakar S, Russell J M, Au W W (1989) Cell-activation responses to cytomegalovirus infection relationship to the phasing of CMV replication and to the induction of cellular damage. Sub-Cellular Biochemistry 15: 157-202.

Almeida-Porada G, Porada C D, Shanley J D, Ascensao J L (1997) Altered production of GM-CSF and IL-8 in cytomegalovirus-infected, IL-1-primed umbilical cord endothelial cells. Exp Hematol 25: 1278-1285.

Almeida-Porada G D, Ascensao J L (1996) Cytomegalovirus as a cause of pancytopenia. Leukemia & Lymphoma 21: 217-223.

Appay V, Brown A, Cribbes S, Randle E, Czaplewski L G (1999) Aggregation of RANTES is responsible for its inflammatory properties. J Biol Chem 274: 27505-27515.

Appay V, Rod-Dunbar P, Cerundolo V, McMichael A, Czaplewski L, Rowland-Jones S (2000) RANTES activates antigen-specific cytotoxic T lymphocytes in a mitogen-like manner through cell surface aggreagation. Internaional Immunology 12: 1173-1182.

Baggiolini M (1998) Chemokines and leukocyte traffic. Nature 392: 565-568.

Baggiolini M, Dewald B, Moser B (1997) Human chemokines: an update. Annu Rev Immunol 15: 675-705.

Baldwin B R, Zhang C O, Keay S (2000) Cloning and epitope mapping of a functional partial fusion receptor for human cytomegalovirus gH. J Gen Virol 81: 27-35.

Bazan J F, Bacon K B, Hardiman G, Wang W, Soo K, Rossi D, Greaves D R, Zlotnik A, Schall T J (1997) A new class of membrane-bound chemokine with a CX3C motif. Nature 385: 640-644.

Beisser P, Grauls G, Bruggeman C, Vink C (1999) Deletion of the R78 G protein-coupled receptor gene from rat cytomegalovirus results in an attenuated, syncytium-inducing mutant strain. J Virol 73: 7218-7230.

Beisser P S, Laurent L, Virelizier J L, Michelson S (2001) Human cytomegalovirus chemokine receptor gene US28 is transcribed in latently infected THP-1 monocytes. J Virol 75: 5949-5957.

Beisser P S, Vink C, Vandam J G, Grauls G, Vanherle S J V, Bruggeman C A (1998) The R33 G Protein-Coupled Receptor Gene Of Rat Cytomegalovirus Plays an Essential Role In the Pathogenesis Of Viral Infection. J Virol 72: 2352-2363.

Billstrom M A, Johnson G L, Avdi N J, Worthen G S (1998) Intracellular signaling by the chemokine receptor US28 during human cytomegalovirus infection. J Virol 72: 5535-5544.

Billstrom M A, Lehman L A, Scott Worthen G (1999) Depletion of extracellular RANTES during human cytomegalovirus infection of endothelial cells. Am J Respir Cell Mol Biol 21: 163-167.

Bodaghi B, Goureau O, Zipeto D, Laurent L, Virelizier J L, Michelson S (1999) Role of IFN-gamma-induced indoleamine 2,3 dioxygenase and inducible nitric oxide synthase in the replication of human cytomegalovirus in retinal pigment epithelial cells. J Immunol 162: 957-964.

Bodaghi B, Jones T R, Zipeto D, Vita C, Sun L, Laurent L, Arenzana-Seisdedos F, Virelizier J L, Michelson S (1998) Chemokine sequestration by viral chemoreceptors as a novel viral escape strategy: withdrawal of chemokines from the environment of cytomegalovirus-infected cells. J Exp Med 188: 855-866.

Bold S, Ohlin M, Garten W, Radsak K (1996) Structural domains involved in human cytomegalovirus glycoprotein B-mediated cell-cell fusion. J Gen Virol 77: 2297-2302.

Boriskin Y S, Moore P, Murday A J, Booth J C, Butcher P D (1999) Human cytomegalovirus genome sequences in lymph nodes. Microbes & Infection 1: 279-283.

Broxmeyer H E, Kim C H (1999) Regulation of hematopoiesis in a sea of chemokine family members with a plethora of redundant activities. Exp Hematol 27: 1113-1123.

Bruening W, Giasson B, Mushynski W, Durham H D (1998) Activation Of Stress-Activated Map Protein Kinases Up-Regulates Expression Of Transgenes Driven By the Cytomegalovirus Immediate/Early Promoter. Nucleic Acids Res 26: 486-489.

Casarosa P, Bakker R, Verzijl D, Navis M, Timmerman H, Smit M (2001) Constitutive siginally of the human cytomegalovirus-encoded chemokine receptor US28. J Biol Chem 276: 1133-1137.

Cashman D, Eaves A, Raines E, Ross R, CJ E (1990) Mechanisma that regulate the cell cycle status of very primitive hematopoietic cells in long-term human marrow cultures. I. Stimulatory role of a variety of mesenchymal cell activators and inhibitory role of TGF-beta. Blood 75: 96-101.

Cha T A, Tom E, Kemble G W, Duke G M, Mocarski E S, Spaete R R (1996) Human Cytomegalovirus Clinical Isolates Carry At Least 19 Genes Not Found In Laboratory Strains. J Virol 70: 78-83.

Chambers J, Angulo A, Amaratunga D, Guo H, Jiang Y, Wan J, Bittner A, Frueh K, Jackson M, Peterson P, Erlander M, Ghazal P (1999) DNA Microarrays of the Complex Human Cytomegalovirus Genome: Profiling Kinetic Class with Drug Sensitivity of Viral Gene Expression. J Virol 73: 5757-5766.

Chee M S, Bankier A T, Beck S, Bohni R, Brown C M, Cerny R, Horsnell T, Hutchinson C A, Kourisarides T, Martignetti J A, Preddi E, Satchwell S C, Tomlinson P, Weston K M, Barrell B G. (1990a). Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. In: "Cytomegaloviruses, Current Topics in

Microbiology and Immunology" (J. M. McDougall, ed.), pp. 125-169. Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, HongKong.

Chee M S, Satchwell S C, Preddie E, Weston K M, Barrell B G (1990b) Human cytomegalovirus encodes three G protein-coupled receptor homologues [see comments]. Nature 344: 774-777.

Choe H, Farzan M, Konkel M, Martin K, Sun Y, Marcon L, Cayabyab M, Berman M, Dorf M E, Gerard N, Gerard C, Sodroski J (1998) The orphan seven-transmembrane receptor apj supports the entry of primary T-cell-line-tropic and dualtropic human immunodeficiency virus type 1. J Virol 72: 6113-6118.

Craigen J L, Grundy J E (1996) Cytomegalovirus Induced Up-Regulation Of Lfa-3 (Cd58) and Icam-1 (Cd54) Is a Direct Viral Effect That Is Not Prevented By Ganciclovir or Foscarnet Treatment. Transplantation 62: 1102-1108.

Craigen J L, Yong K L, Jordan N J, MacCormac L P, Westwick J, Akbar A N, Grundy J E (1997) Human cytomegalovirus infection up-regulates interleukin-8 gene expression and stimulates neutrophil transendothelial migration. Immunology 92: 138-145.

Crapnell K, Zanjani E D, Chaudhuri A, Ascensao J L, St Jeor S, Maciejewski J P (2000) In vitro infection of megakaryocytes and their precursors by human cytomegalovirus. Blood 95: 487-493.

Davis-Poynter N J, Lynch D M, Vally H, Shellam G R, Rawlinson W D, Barrell B G, Farrell H E (1997) Identification and Characterization Of a G Protein-Coupled Receptor Homolog Encoded By Murine Cytomegalovirus. J Virol 71: 1521-1529.

Davis-Poynter N J, Rawlinson W D, Barrell B G, Shellam G R, Farrell H E (1995) Identification and Characterisation of a G-protein Coupled Receptor Homologue Encoded by Murine Cytomegalovirus. The 20th International Herpesvirus Workshop, Groningen, University of Groningen, Program & Abstracts: 88 (Abstract).

Diasbaruffi M, Pereiradasilva G, Jamur M C, Roquebarreira M C (1998) Heparin Potentiates In Vivo Neutrophil Migration Induced By Il-8. Glycoconjugate Journal 15: 523-526.

Faure M, Voyno-Yasenetskaya T A, Bourne H R (1994) cAMP and beta subunits of heterotrimeric G proteins stimulate the mitogen-activated protein kinase pathway in COS-7 cells.

Fest T, Angonin R, Mougin C, Deschaseaux M, Lab M, Cahn J Y, Herve P (1994a) Detection of cytomegalovirus-infected cells in bone marrow biopsy specimens obtained before allogeneic bone marrow transplantation from donors and recipients. Transplantation 57: 1681-1683.

Fest T, Deschaseaux M, Mougin C, Cahn J Y, Dupond J L, Herve P (1994b) In situ hybridization for the detection of cytomegalovirus in blood or bone marrow leucocytes after allogeneic bone marrow transplantation. Br J Haematol 86: 619-623.

Fietze E, Prösch S, Reinke P, al e (1994) Cytomegalovirus infection in transplant receipients. The role of tumor necrosis factor. Transplantation 58: 675-680.

Fraile-Ramos A, Kledal T N, Pelchen-Matthews A, Bowers K, Schwartz T W, Marsh M (2001) The human cytomegalovirus us28 protein is located in endocytic vesicles and undergoes constitutive endocytosis and recycling. Mol Biol Cell 12: 1737-1749.

Gao J L, Kuhns D B, Tiffany H L, McDermott D, Li X, Francke U, Murphy P M (1993) Structure and functional expression of the human macrophage inflammatory protein 1 alpha/RANTES receptor. J Exp Med 177: 1421-1427.

Gao J L, Murphy P M (1994) Human cytomegalovirus open reading frame US28 encodes a functional beta chemokine receptor. J Biol Chem 269: 28539-28542.

Garnett H M (1979) The early effects of human cytomegalovirus infection on macromolecular synthesis in human embryonic fibroblasts. Brief report. Arch Virol 60: 147-151.

Gerna G, Percivalle E, Baldanti F, Sozzani S, Lanzarini P, Genini E, Lilleri D, Revello M G (2000) Human Cytomegalovirus Replicates Abortively in Polymorphonuclear Leukocytes after Transfer from Infected Endothelial Cells via Transient Microfusion Events. J Virol 74: 5629-5638.

Goh C-S, Bogan A A, Joachimiak M, Walther D,. Cohen F E (2000) Co-evolution of proteins with their interaction partners. J Mol Biol **299:** 283-293.

Gompels U A, Nicholas J, Lawrence G, Jones M, Thomson B J, Martin M E, Efstathiou S, Craxton M, Macaulay H A (1995) The DNA sequence of human herpesvirus-6: structure, coding content, and genome evolution. Virology 209: 29-51.

Grundy J E, Lawson K M, MacCormac L P, Fletcher J M, Yong K L (1998) Cytomegalovirus-infected endothelial cells recruit neutrophils by the secretion of C-X-C chemokines and transmit virus by direct neutrophil-endothelial cell contact and during neutrophil transendothelial migration. J Infect Dis 177: 1465-1474.

Gutkind J S (1998) THE PATHWAYS CONNECTING G PROTEIN-COUPLED RECEPTORS TO THE NUCLEUS THROUGH DIVERGENT MITOGEN-ACTIVATED PROTEIN KINASE CASCADES [Review]. J Biol Chem 273: 1839-1842.

Hadida F, Vieillard V, Autran B, Lewis-Clark I, Baggiolini M, Debré P (1998) HIV-specific T cell cytotoxicity mediated by RANTES via the chemokine receptor CCR3. J Exp Med 188: 609-614.

Hahn G, Jores R, Mocarski E S (1998) Cytomegalovirus remains latent in a common precursor of dendritc and myeloid cells. Proc. Nalt. Acad. Sci (USA) 95: 3937-3942.

Hahn G, Mocarski E. (1996). Human cytomegalovirus latency and latently-infected cell types in hematopoeitic progenitors and peripheral blood (Abs. N° 303). In "21st Herpesvirus Workshop". Northern Illinois University, DeKalb, Ill. USA.

Hamm H E (1998) THE MANY FACES OF G PROTEIN SIGNALING [Review]. J Biol Chem 273: 669-672.

Haskell C A, Cleary M D, Charo I F (2000) Unique role of the chemokine domain of fractalkine in cell capture: Kinetics of receptor dissociation correlate with cell adhesion. J Biol Chem 275: 34183-34189.

Hendrix R M, Wagenaar M, Slobbe R L, Bruggeman C A (1997) Widespread presence of cytomegalovirus DNA in tissues of healthy trauma victims. J Clin Pathol 50: 59-63.

Hirsch A J, Shenk T (1999) Human cytomegalovirus inhibits transcription of the CC chemokine MCP-1 gene. J Virol 73: 404-410.

Humar A, St Louis P, Mazzulli T, McGeer A, Lipton J, Messner H, MacDonald K S (1999) Elevated serum cytokines are associated with cytomegalovirus infection and disease in bone marrow transplant recipients. J Infect Dis 179: 484-488.

Imai T, Hieshima K, Haskell C, Baba M, Nagira M, Nishimura M, Kakizaki M, Takagi S, Nomiyama H, Schall T J, Yoshie O (1997) Identification and Molecular Characterization Of Fractalkine Receptor Cx(3)Cr1, Which Mediates Both Leukocyte Migration and Adhesion. Cell 91: 521-530.

Isegawa Y, Ping Z, Nakano K, Sugimoto N, Yamanishi K (1998) Human herpesvirus 6 open reading frame U12 encodes a functional beta-chemokine receptor. J Virol 72: 6104-6112.

Kledal T N, Rosenkilde M M, Schwartz T W (1998) Selective recognition of the membrane-bound CX3C chemokine, fractalkine, by the human cytomegalovirus-encoded broad-spectrum receptor US28. FEBS Lett 441: 209-214.

Kondo K, Kaneshima H, Mocarski E S (1994) Human cytomegalovirus latent infection of granulocyte-macrophage progenitors. Proc Natl Acad Sci USA 91: 11879-11883.

Kondo K, Mocarski E S (1995) Cytomegalovirus Latency and Latency-Specific Transcription In Hematopoietic Progenitors. Scand J Infect Dis: 63-67.

Kuhn D E, Beall C J, Kolattukudy P E (1995) The cytomegalovirus US28 protein binds multiple CC chemokines with high affinity. Biochem Biophys Res Commun 211: 325-330.

Lagneaux L, Delforge A, Bron D, Bosmans E, Stryckmans P (1995) Comparative analysis of cytokines released by bone marrow stromal cells from normal donors and B-cell chronic lymphocytic leukemic patients. Leuk Lymphoma 17: 127-133.

Lagneaux L, Delforge A, Snoek R, Bosmans E, Schols D, De Clercq E, Stryckmans P, Bron D (1996) Imbalance in production of cytokines by bone marrow stromal cells folowing cytomegalovirus infection. J Infect Dis 174: 913-919.

Larsson S, Soderberg-Naucler C, Wang F Z, Moller E (1998) Cytomegalovirus DNA can be detected in peripheral blood mononuclear cells from all seropositive and most seronegative healthy blood donors over time. Transfusion 38: 271-278.

Locati M, Lamorte G, Luini W, Introna M, Bernasconi S, Mantovani A, Sozzani S (1996) Inhibition Of Monocyte Chemotaxis to C-C Chemokines By Antisense Oligonucleotide For Cytosolic Phospholipase a(2). J Biol Chem 271: 6010-6016.

Loetscher P, Moser B, Baggiolini M (2000) Chemokines and their receptors in lymphocyte traffic and HIV infection. Adv Immunol 74: 127-180.

Luster A D, Greenberg S M, Leder P (1995) The IP-10 chemokine binds to a specific cell surface heparan sulfate site shared with platelet factor 4 and inhibits endothelial cell proliferation. J Exp Med 182: 219-231.

Margulies B J, Browne H, Gibson W (1996) Identification of the human cytomegalovirus G protein-coupled receptor homologue encoded by UL33 in infected cells and enveloped virus particles. Virology 225: 111-125.

Mendelson M, Monard S, Sissons P, Sinclair J (1996) Detection Of Endogenous Human Cytomegalovirus In Cd34(+) Bone Marrow Progenitors. J Gen Virol 77: 3099-3102.

Menotti L, Mirandola P, Locati M, Campadelli-Fiume G (1999) Trafficking to the plasma memvrane of the seven-transmembrane protein encoded by human

herpesvirus 6 U51 gene involves a cell-specific function present in T lymphocytes. J Virol 73: 325-333.

Michelson S (1997) Interaction of human cytomegalovirus with monocytes/macrophages: a love-hate relationship. Pathol Biol (Paris) 45: 146-158.

Michelson S, Dal Monte P, Zipeto D, Bodaghi B, Laurent L, Oberlin E, Arenzana-Seisdedos F, Virelizier J L, Landini M P (1997) Modulation of RANTES production by human cytomegalovirus infection of fibroblasts. J Virol 71: 6495-6500.

Milne R B S, Mattick C, Nicholson L, Alcami A, Gompels U A (2000) RANTES binding and down-regulation by a novel human herpesvirus-6 chemokine receptor. J Immunol 164: 2396-2404.

Minton E J, Tysoe C, Sinclair J H, Sissons J G (1994) Human cytomegalovirus infection of the monocyte/macrophage lineage in bone marrow. J Virol 68: 4017-4021.

Mizoue L S, Sullivan S K, King D S, Kledal T N, Schwartz T W, Bacon K B, Handel T M (2001) Molecular determinants of receptor binding and signaling by the CX3C chemokine fractalkine. J Biol Chem 29: 29.

Monti G, Magnan A, Fattal M, Rain B, Humbert M, Mege J L, Noirclerc M, Dartevelle P, Cerrina J, Simonneau G, Galanaud P, Emilie D (1996) Intrapulmonary production of RANTES during rejection and CMV pneumonitis after lung transplantation. Transplantation 61: 1757-1762.

Murayama T, Ohara Y, Obuchi M, Khabar K S, Higashi H, Mukaida N, Matsushima K (1997) Human cytomegalovirus induces interleukin-8 production by a human monocytic cell line, THP-1, through acting concurrently on AP-1- and NF-kappaB-binding sites of the interleukin-8 gene. J Virol 71: 5692-5695.

Murphy P M (2000) Viral antichemokines: From pathogenesis to drug discovery. J Clin Invest 105: 1515-1517.

Murphy P M, Baggiolini M, Charo I F, Herbert C A, Horuk, R;, Matsushima K, Miller L H, Oppenheim J J, Power, C.A. (2000) International Union of Pharmacology. XXII. Nomenclature for Chemokine Receptors. Pharmacological Reviews 52: 145-176.

Myerson D, Hackman R C, Nelson J A, Ward D C, McDougall J K (1984) Widespread presence of histologically occult cytomegalovirus. Hum Pathol 15: 430-439.

Navarro D, Paz P, Tugizov S, Topp K, La Vail J, Pereira L (1993) Glycoprotein B of human cytomegalovirus promotes virion penetration into cells, transmission of infection from cell to cell, and fusion of infected cells. Virology 197: 143-158.

Neote K, DiGregorio D, Mak J Y, Horuk R, Schall T J (1993) Molecular cloning, functional expression, and signaling characteristics of a C-C chemokine receptor. Cell 72: 415-425.

Nordoy I, Muller F, Nordal K P, Rollag H, Lien E, Aukrust P, Froland S S (1999) Immunologic parameters as predictive factors of cytomegalovirus disease in renal allograft recipients. J Infect Dis 180: 195-198.

Ohagen A, Li L, Rosenzweig A, Gabuzda D (2000) Cell-dependent mechanisms restrict the HIV type 1 coreceptor activity of US28, a chemokine receptor homolog encoded by human cytomegalovirus. AIDS Res Hum Retroviruses 16: 27-35.

Oravecz T, Pall M, Wang J, Roderiquez G, Ditto M, Norcross M A (1997) Regulation of anti-HIV-1 activity of RANTES by heparan sulfate proteoglycans. J Immunol 159: 4587-4592.

Pablos J L, Amara A, Bouloc A, Santiago B, Caruz A, Galindo M, Delaunay T, Virelizier J L, Arenzana-Seisdedos F (1999) Stromal-cell derived factor is expressed by dendritic cells and endothelium in human skin. Am J Pathol 155: 1577-1586.

Patterson B K, Landay A, Andersson J, Brown C, Behbahani H, Jiyamapa D, Burki Z, Stanislawski D, Czerniewski M A, Garcia P (1998) Repertoire of chemokine receptor expression in the female genital tract: implications for human immunodeficiency virus transmission. Am J Pathol 153: 481-490.

Penchansky L, Krause J R (1979) Identification of cytomegalovirus in bone marrow biopsy. Southern Medical Journal 72: 500-501.

Penfold M E, Dairaghi D J, Duke G M, Saederup N, Mocarski E S, Kemble G W, Schall T J (1999) Cytomegalovirus encodes a potent alpha chemokine. Proc Natl Acad Sci U S A 96: 9839-9844.

Pleskoff O, Treboute C, Alizon M (1998) The cytomegalovirus-encoded chemokine receptor US28 can enhance cell-cell fusion mediated by different viral proteins. J Virol 72: 6389-6397.

Pleskoff O, Treboute C, Brelot A, Heveker N, Seman M, Alizon M (1997) Identification of a chemokine receptor encoded by human cytomegalovirus as a cofactor for HIV-1 entry [see comments]. Science 276: 1874-1878.

Randolph-Habecker J, Beall C J, Kolattukudy P E, Sedmak D D (1997) Monocyte chemoattractant protein-1 binding by cytomegalovirus-infected endothelial cells. Transplant Proc 29: 807-808.

Reschke M, Revello M G, Percivalle E, Radsak K, Landini M P (1999) Constitutive expression of human cytomegalovirus (HCMV) glycoprotein gpUL75 (gH) in astrocytoma cells: a study of the specific humoral immune response. Viral Immunol 12: 249-262.

Revello M G, Percivalle E, Arbustini E, Pardi R, Sozzani S, Gerna G (1998) In Vitro Generation Of Human Cytomegalovirus Pp65 Antigenemia, Viremia, and Leukodnaemia. J Clin Invest 101: 2686-2692.

Rodems S M, Spector D H (1998) Extracellular signal-regulated kinase activity is sustained early during human cytomegalovirus infection. J Virol 72: 9173-9180.

Rucker J, Edinger A L, Sharron M, Samson M, Lee B, Berson J F, Yi Y, Margulies B, Collman R G, Doranz B J, Parmentier M, Doms R W (1997) Utilization of chemokine receptors, orphan receptors, and herpesvirus-encoded receptors by diverse human and simian immunodeficiency viruses. J Virol 71: 8999-9007.

Saederup N, Lin Y C, Dairaghi D J, Schall T J, Mocarski E S (1999) Cytomegalovirus-encoded beta chemokine promotes monocyte-associated viremia in the host. Proc Natl Acad Sci U S A 96: 10881-10886.

Schaarschmidt P, Reinhardt B, Michel D, Vaida B, Mayr K, Luske A, Baur R, Gschwend J, Kleinschmidt K, Kountidis M, Wenderoth U, Voisard R, Mertens T (1999) Altered Expression of Extracellular Matrix in Human-Cytomegalovirus-Infected Cells and a Human Artery Organ Culture Model to Study Its Biological Relevance. Intervirology 42: 365-372.

Selbie L A, Hill S J (1998) G protein-coupled-receptor cross-talk: the fine-tuning of multiple receptor-signalling pathways. Trends Pharmacol Sci 19: 87-93.

Shibutani T, Johnson T M, Yu Z X, Ferrans V J, Moss J, Epstein S E (1997) Pertussis Toxin-Sensitive G Proteins As Mediators Of the Signal Transduction Pathways Activated By Cytomegalovirus Infection Of Smooth Muscle Cells. J Clin Invest 100: 2054-2061.

Signoret N, Marsh M. Analysis of chemokine receptor endocytosis and recycling. In: "Chemokine Protocols". Humana Press, Inc., Totowa, New Jersey, USA.

Sindre H, Rollag H, Degre M, Hestdal K (2000) Human cytomegalovirus induced inhibition of hematopoietic cell line growth is initiated by events taking place before translation of viral gene products. Arch Virol 145: 99-111.

Sinzger C, Jahn G (1996) Human cytomegalovirus cell tropism and pathogenesis. Intervirology 39: 302-319.

Sinzger C, Knapp J, Plachter B, Schmidt K, Jahn G (1997) Quantification Of Replication Of Clinical Cytomegalovirus Isolates In Cultured Endothelial Cells and Fibroblasts By a Focus Expansion Assay. J Virol Methods 63: 103-112.

Sinzger C, Plachter B, Grefte A, The T H, Jahn G (1996) Tissue Macrophages Are Infected By Human Cytomegalovirus In Vivo. J Infect Dis 173: 240-245.

Soderberg-Naucler C, Fish K N, Nelson J A (1997) Reactivation Of Latent Human Cytomegalovirus By Allogeneic Stimulation Of Blood Cells From Healthy Donors. Cell 91: 119-126.

Speir E, Shibutani T, Yu Z X, Ferrans V, Epstein S E (1996) Role Of Reactive Oxygen Intermediates In Cytomegalovirus Gene Expression and In the Response Of Human Smooth Muscle Cells to Viral Infection. Circ Res 79: 1143-1152.

Speir E, Yu Z X, Ferrans V J, Huang E S, Epstein S E (1998) Aspirin Attenuates Cytomegalovirus Infectivity and Gene Expression Mediated By Cyclooxygenase-2 In Coronary Artery Smooth Muscle Cells. Circ Res 83: 210-216.

Streblow D N, Soderberg-Naucler C, Vieira J, Smith P, Wakabayashi E, Ruchti F, Mattison K, Altschuler Y, Nelson J A (1999) The human cytomegalovirus chemokine receptor US28 mediates vascular smooth muscle cell migration. Cell 99: 511-520.

Toorkey C B, Carrigan D R (1989) Immunohistochemical detection of an immediate early antigen of human cytomegalovirus in normal tissues. J Infect Dis 160: 741-751.

Vieira J, Schall T J, Corey L, Geballe A P (1998) Functional analysis of the human cytomegalovirus US28 gene by insertion mutagenesis with the green fluorescent protein gene. J Virol 72: 8158-8165.

Vonlaer D, Meyerkoenig U, Serr A, Finke J, Kanz L, Fauser A A, Neumannhaefelin D, Brugger W, Hufert F T (1995) Detection Of Cytomegalovirus Dna In Cd34(+) Cells From Blood and Bone Marrow. Blood 86: 4086-4090.

Welch A R, McGregor L M, Gibson W (1991) Cytomegalovirus homologs of cellular G protein-coupled receptor genes are transcribed. J Virol 65: 3915-3918.

Yurochko A D, Huang E S (1999) Human cytomegalovirus binding to human monocytes induces immunoregulatory gene expression. J Immunol 162: 4806-4816.

Zhu H, Cong J P, Shenk T (1997) Use Of Differential Display Analysis to Assess the Effect Of Human Cytomegalovirus Infection On the Accumulation Of Cellular Rnas

UC
Franc
LIBRAR

UC
Franc
LIBRAR

UC
Fra
LIBRA

- Induction Of Interferon-Responsive Rnas. Proc Natl Acad Sci USA 94: 13985-13990.

Zhuravskaya T, Maciejewski J, Massey R, St. Jeor S. (1996). Human cytomegalovirus (HCMV) infection of hematopoietic progenitor cells (Abs N° 301). In "21st Herpesvirus Workshop", Northern Illinos University, DeKalb, Ill. USA.

Zipeto D, Bodaghi B, Laurent L, Virelizier J L, Michelson S (1999) Kinetics of transcription of human cytomegalovirus chemokine receptor US28 in different cell types. J Gen Virol 80: 543-547.

Chapter 3

Molecular Phylogeny and Evolution of the Plant-Specific Seven Transmembrane MLO Family

This chapter is in press as:

Devoto A, Hartmann HA, Piffanelli P, Elliott C, Simmons C, Taramino G, Goh CS, Cohen FE, Emerson BC, Schulze-Lefert P, and Panstruga R. Molecular phylogeny and evolution of the plant-specific seven transmembrane MLO family. *J Mol Evol.*

Abstract

Homologs of barley *Mlo* encode the only family of seven transmembrane (TM) proteins in plants. Their topology, subcellular localization, and sequence diversification is highly reminiscent of G-protein coupled receptors (GPCRs) from animals and fungi. We present a computational analysis of MLO family members based on 31 full-size and three partial sequences, which originate from several monocot species, the dicot *Arabidopsis thaliana*, and the moss *Ceratodon purpureus*. This enabled us to date back the origin of the *Mlo* gene family at least to the early stages of land plant evolution. Genomic organization of the corresponding genes supports a monophyletic origin of the *Mlo* gene family. Phylogenetic analysis revealed five clades of which three contain both monocot and dicot members whilst two indicate class-specific diversification. Analysis of the ratio of non-synonymous and synonymous changes in coding sequences provided evidence for functional constraint on the evolution of the DNA sequences and purifying selection, which appears to be reduced in the first extracellular loop of twelve closely related orthologs. The 31 full-size sequences were examined for potential domain-specific intramolecular co-evolution. This revealed evidence for concerted evolution of all three cytoplasmic domains with each other and the C-terminal cytoplasmic tail, suggesting interplay of all intracellular domains for MLO function.

Keywords**: 7 TM protein, co-evolution, gene family, exon/exon junctions, Mlo, G-protein coupled receptor**

Running head: **MLO phylogeny and evolution**

Abbreviations: **EST, expressed sequence tag; GPCR, G-protein coupled receptor; TM, transmembrane**

Abstract

Homologs of barley *Mlo* encode the only family of seven transmembrane (TM) proteins in plants. Their topology, subcellular localization, and sequence diversification is reminiscent of G-protein coupled receptors (GPCRs) from animals and fungi. We present a computational analysis of MLO family members based on 31 full-size and three partial sequences, which originate from several monocot species, the dicot *Arabidopsis thaliana*, and the moss *Ceratodon purpureus*. This enabled us to date back the origin of the *Mlo* gene family at least to the early stages of land plant evolution. Genomic organization of the corresponding genes supports a monophyletic origin of the *Mlo* gene family. Phylogenetic analysis revealed five clades of which three contain both monocot and dicot members whilst two indicate class-specific diversification. Analysis of the ratio of non-synonymous and synonymous changes in coding sequences provided evidence for functional constraint on the evolution of the DNA sequences and purifying selection, which appears to be reduced in the first extracellular loop of twelve closely related orthologs. The 31 full-size sequences were examined for potential domain-specific intramolecular co-evolution. This revealed evidence for concerted evolution of all three cytoplasmic domains with each other and the C-terminal cytoplasmic tail, suggesting interplay of all intracellular domains for MLO function.

Introduction

In barley, presence of the wild-type *Mlo* gene modulates defense responses to the powdery mildew fungus, *Blumeria graminis* f sp *hordei* (Büschges et al. 1997). Homozygous *mlo* mutant plants exhibit full resistance to the fungal pathogen whereas *Mlo* overexpression results in super-susceptibility (Wolter et al. 1993, Kim et al. 2002b). MLO is likely to have a role in additional biological processes since axenically grown *mlo* mutant plants show accelerated leaf senescence symptoms and a spontaneous cell death phenotype (Wolter et al. 1993, Peterhänsel et al. 1997, Piffanelli et al. in press). This suggests a function for MLO in cell death protection upon biotic stress and leaf senescence. Two genes, *Ror1* and *Ror2*, have been described that are required for full *mlo*-mediated resistance. Mutations in either of these genes confer partial susceptibility in an *mlo* mutant background and also compromise the spontaneous cell death phenotype (Freialdenhoven et al. 1996, Peterhänsel et al. 1997).

To date, MLO is the only plant polytopic membrane protein experimentally shown to consist of seven membrane-spanning domains (Devoto et al. 1999). However, a further protein, the putative GPCR GCR1, is predicted to also contain seven TM helices (Josefsson and Rask, 1997, Plakidou-Dymock et al. 1998). The barley MLO protein resides in the plasma membrane, with the N-terminus positioned extracellularly and the C-terminus intracellularly (Devoto et al. 1999). Database searches have revealed that MLO belongs to a gene family that is restricted to the plant kingdom. Inspection of the near full-length Arabidopsis genome has shown that *Mlo*-like genes represent the only sequence-diversified family encoding 7TM proteins in plants whilst *GCR1* is a single

copy gene (Devoto et al. 1999, The Arabidopsis Genome Initiative, 2000). To date, all known animal and fungal (including yeast) sequence-diversified protein families with a 7TM topology function as GPCRs, which relay extracellular signals into an intracellular response by activating a heterotrimeric G-protein (Bockaert and Pin, 1999). Recent data, however, indicate that MLO-mediated defence suppression in barley functions independently of heterotrimeric G-proteins and that calmodulin interacts with MLO to dampen defence reactions against the powdery mildew fungus (Kim et al., 2002b).

Here we present a thorough computational analysis of the MLO protein family based on a comprehensive set of sequences derived from Arabidopsis and maize to trace back the phylogenetic history of these plant-specific proteins. We have investigated the data set for the presence of domain-specific adaptive molecular evolution. A recently developed algorithm that allows the identification of protein-protein interaction pairs identified candidate domains that have evolved in a concerted manner. Our findings are consistent with a presumptive receptor function of MLO proteins.

Materials and Methods

Mlo DNA sequences

Mlo cDNA sequences from Arabidopsis were obtained by reverse transcriptase polymerase chain reaction using oligonucleotides that were derived from the publicly available genomic sequences. Similarly, cDNAs of *TaMlo1*, *TaMlo2*, and *OsMlo2* and genomic sequences of *Mlo2* and *OsMlo1* were obtained using standard procedures (details about the isolation of these clones will be published elsewhere). Sequence information about *Zea mays Mlo* cDNAs (*ZmMlo1-9*) were derived from corresponding expressed sequence tag (EST) clones from the combined DuPont/Pioneer EST collection. Nucleotide sequences of all cDNAs were determined by applying standard techniques on ABI373/377 automated sequencers.

Phylogenetic analyses

Protein sequences were aligned using PileUp (Wisconsin Package Version 10.0, Genetics Computer Group, Madison, WI, USA) and optimized by hand. Phylogenetic analyses were performed using the maximum parsimony search optimality criterion of PAUP* v.4.0b8 (Swofford, 1998). Maximum parsimony analysis of protein sequences was performed for (i) full length sequences excluding N- and C-termini, (ii) all transmembrane regions only, (iii) all extracellular and intracellular regions, (iv) all extracellular regions, and (v) all intracellular regions. An additional analysis was performed for a partial sequence alignment including an MLO homolog of a moss, *Ceratodon purpureus*. Searches were performed using the heuristic search option and all

trees were rooted using the mid-point rooting option. Support for the branching arrangements was evaluated by bootstrap analyses using 1000 replicates.

Calculating d_N/d_S ratios

To calculate the ratio of non-synonymous to synonymous substitutions (d_N/d_S) we used the yn00 program of PAML (Yang 1997) implementing the method of Yang and Nielsen (2000). For these analyses we used an alignment of one wheat (*TaMlo2*) sequence and 11 sequences derived from nine different species of the genus *Hordeum*. The *Hordeum* sequences correspond to amino acid residues 69-145 of barley MLO, covering the first extracellular loop and some neighboring residues, and were obtained by standard PCR amplification using genomic DNA as template and oligonucleotides Mlo4 5′-AAGGCGGAGCTCATGCTGGTGGGC-3′ and Mlo5 5′-ACGGCTTAGAGCTATGGTGATGAC-3′ as primers. Amplification products (~350-400 bp, including one intron) were purified on agarose gels, subcloned in pGEM-Teasy (Promega) and subjected to sequence analysis. We dissected the resulting nucleotide sequences (excluding primer and intron sequences) into three parts that were investigated separately; (i) the whole stretch, corresponding to amino acids 69-145 of barley MLO (ii) extracellular loop 1 excluding the region between conserved cysteine residues 86 and 114, and (iii) the region between conserved cysteine residues 86 and 114. The yn00 program calculates d_N/d_S ratios for each pairwise comparison. We have then summarised these as an average d_N/d_S ratio for each region (excluding ratios that had a zero value for either d_N or d_S) to compare differences on the rate of amino acid substitution among the three regions.

Co-evolution analysis

The correlation analysis was done on every possible domain-domain pair using methods described previously (Goh et al. 2000). Distance matrices were generated from the multiple alignments using ClustalW (Thompson et al. 1994). We employed a linear regression analysis measuring the c orrelation between pair wise evolutionary distances among all peptides in a multiple sequence alignment. These were correlated with the evolutionary distances among the corresponding binding partners using the linear correlation coefficient *r* (Pearson's correlation coefficient (Press et al. 1998) between the distance matrices of all possible interacting domains where $-1 \leq r \leq +1$. Positive values of *r* would indicate a positive correlation, and *r*-values of around zero would indicate no correlation. Additionally, negative values of *r* would indicate anti-correlation.

Results and Discussion

Phylogenetic analysis of Mlo-like genes suggests an origin in the early stages of land plant evolution

Previously, we described the existence of *Mlo*-like sequences in different monocot and dicot species (Devoto et al. 1999). In the meantime, further genomic sequences and ESTs sequence-related to barley *Mlo* were released. By searching the public databases using the BLAST or PSI-BLAST algorithms (http://www.ncbi.nlm.nih.gov/BLAST/), *Mlo*-like genes were identified in an even broader range of monocotyledonous (*Hordeum vulgare, Oryza sativa, Secale cereale, Triticum aestivum, Zea mays,*) as well as dicotyledonous plant species (*Arabidopsis thaliana*, *Brassica rapa*, *Citrullus lanatus*, *Glycine max*, *Gossypium hirsutum, Linum usatissimum, Lotus japonicus*, *Lycopersicon esculentum*, *Medicago truncatula, Solanum tuberosum*, *Sorghum bicolor*). Multiple distinct genes were found in most of these species, indicating their organization into multigene families. Recently, the nearly full genomic sequence of *Arabidopsis thaliana* was released (The Arabidopsis Genome Initiative 2000), covering more than 90% of the 125 Mb genome of the weed. Based on this data, we identified 15 distinct members for which full-length genomic sequences are known (Table 3.1). The remaining 10 Mb of the Arabidopsis genomic sequence are supposed to cover mainly rDNA repeat units, centromeric and telomeric regions as well as other regions of complex sequence structure that are unlikely to harbor many coding sequences. Thus, we conclude that the 15 Arabidopsis *Mlo* homologs identified to date are likely to represent the actual number. A former estimate

of 25-35 homologs (Devoto et al. 1999) is apparently due to an overrepresentation of *Mlo* homologs in early released sequences of the Arabidopsis genome. The designation of the 15 genes is given in Table 3.1 (see also http://www.arabidopsis.org/info/genefamily/mlo.html). Only eight of these are currently represented by corresponding ESTs in GenBank, indicating their generally low expression levels. However, we were able to isolate matching cDNAs for all members by reverse transcriptase polymerase chain reaction. Subsequent DNA sequencing confirmed the identity of the clones, demonstrating that all 15 members are expressed, albeit at low levels (Table 3.1 and data not shown).

To identify *Mlo* family members in the monocotyledonous plant *Zea mays*, we searched the Pioneer/DuPont maize EST database which to date comprises 400,000 ESTs. Nucleotide sequences of nine distinct *Mlo* genes were identified in this database (seven of which appeared to be full-length), indicative of a similar total number of *Mlo* genes in maize and Arabidopsis. Like Arabidopsis, most of the maize genes are expressed either at a low level or preferentially in particular tissues (data not shown).

Except for barley *Mlo*, no biological function has been assigned to any other *Mlo*-like gene to date. We have isolated cDNAs from wheat and rice that are exceptionally similar to barley *Mlo*. Due to their syntenic genomic locations relative to the barley gene on chromosome 4H, these members are likely to be orthologs. In single-cell transfection experiments of barley *mlo* mutants (Shirasu et al. 1999), *OsMlo2* and *TaMlo2* showed either full (*TaMlo2*) or partial (*OsMlo2*) complementation, indicating that during evolution the function of these orthologs were preserved (Elliott et al. in press). A comprehensive list of all 34 members analyzed here is shown in Table 3.1.

Table 3.1. Compilation of *Mlo* homologs

Gene	Organism	GenBank (cDNA)	GenBank (genomic)	Genome position	Introns	Amino acids
AtMlo1	*A. thaliana*	Z95352	At4g02600	Chr.IV, 15 cM	11	526
AtMlo2	*A. thaliana*	AF369563	At1g11310	Chr.I, 10 cM	13	573
AtMlo3	*A. thaliana*	AF369564	At3g45290	Chr.III, 61 cM	14	508
AtMlo4	*A. thaliana*	AF369565	At1g11000	Chr.I, 10 cM	14	570
AtMlo5	*A. thaliana*	AF369566	At2g33670	Chr.II, 76 cM	14	500
AtMlo6	*A. thaliana*	AF369567	At1g61560	Chr.I, 84 cM	13	583
AtMlo7	*A. thaliana*	AF369568	At2g17430	Chr.II, 32 cM	13	542
AtMlo8	*A. thaliana*	AF369569	At2g17480	Chr.II, 32 cM	14	593
AtMlo9	*A. thaliana*	AF369570	At1g42560	Chr.I, 62 cM	14	460
AtMlo10	*A. thaliana*	AF369571	At5g65970	Chr.V, 128 cM	14	569
AtMlo11	*A. thaliana*	AF369572	At5g53760	Chr.V, 100 cM	14	565
AtMlo12[a]	*A. thaliana*	AF369573	At2g39200	Chr.II, 72 cM	14	576

AtMlo13[b]	*A. thaliana*	AF369574	At4g24250	Chr.IV, 83 cM	13	478
AtMlo14	*A. thaliana*	AF369575	At1g26700	Chr.I, 38 cM	14	550
AtMlo15	*A. thaliana*	AF369576	At2g44110	Chr.II, 78 cM	13	496
CpMlo	*C. purpureus*	AW087034	--	n.d.	--	p.s.
Mlo	*H. vulgare*	Z83834	Y14573	Chr. IV	11	533
Mlo2	*H. vulgare*	--	Z95496	Chr. IV	11	544
OsMlo1	*O. sativa*	--	Z95353	Chr.VI	12	540
OsMlo2	*O. sativa*	AF384030	AP000615	Chr.III	12	555
OsMlo3	*O. sativa*	AF388195	--	n.d.	--	554
OsMlo4	*O. sativa*	--	AC073166	Chr. X	14	580
TaMlo1	*T.aestivum*	AX063298	--	n.d.	--	534
TaMlo2	*T. aestivum*	AX063294	--	n.d.	--	534
TaMlo3	*T. aestivum*	AX063296		n.d.	--	534
ZmMlo1	*Z. mays*	AY029312	--	Chr.I, bin 1	--	563
ZmMlo2	*Z. mays*	AY029313	--	Chr.I, bin 4	--	565

ZmMlo3	*Z. mays*	AY029314	--	Chr.II, bin 4	--	496
ZmMlo3	*Z. mays*	AY029314	--	Chr.II, bin 4	--	496
ZmMlo4	*Z. mays*	AY029315	--	Chr.III, bin 5	--	509
ZmMlo5	*Z. mays*	AY029316	--	Chr.III, bin 6	--	p.s.
ZmMlo6	*Z. mays*	AY029317	--	Chr.V, bin 4/5	--	515
ZmMlo7	*Z. mays*	AY029318	--	Chr.IX, bin 4	--	499
ZmMlo8	*Z. mays*	AY029319	--	Chr.VI, bin 5-7	--	492
ZmMlo9	*Z. mays*	AY029320	--	n.d.	--	p.s.

p.s.; partial sequence

--; genomic or cDNA sequence not available

[a] formerly designated as *AtMlo18* (Devoto *et al.* 1999)

[b] formerly designated as *AtMlo20* (Devoto *et al.* 1999)

Phylogenetic analysis performed on 31 MLO full-length protein sequences identifies six subfamilies comprised of five clades (I–V), with strong bootstrap support for the monophyly of each clade, and a single divergent lineage (AtML03; Fig. 3.1). There is also strong bootstrap support for a sister group relationship between subfamilies I and II, while relationships among the remaining subfamilies are unresolved. With a few exceptions, phylogenetic analyses of specific regions of the *Mlo* genes also recover these six subfamilies with moderate to high bootstrap support (Table 3.2). On average, subfamily members exhibit 45% identity and 70% similarity at the amino acid level. Interestingly, subfamily IV comprises only monocot homologs, including the presumptive orthologs from barley, wheat, and rice. Similarly, three Arabidopsis members (AtMLO2, AtMLO6 and AtMLO12) cluster together and define subfamily V, which appears to be restricted to dicots (or, alternatively, to Arabidopsis) given the fact that the analysis of 400,000 maize ESTs failed to reveal members of this gene cluster.

The results of the phylogenetic analysis support an early evolutionary diversification of the MLO subgroups, well before the origin of monocots and dicots. MLO homologs of *Arabidopsis* and *Zea mays* are highly divergent with representatives in clades I, II, III, V and clades I, II, III, IV, respectively. Maintenance of these subfamilies (clades) may indicate preservation of an early functional diversification. Whether monocot- and dicot-specific clades IV and V emerged after the separation of these two classes or whether members of these clades were lost subsequently remains elusive.

Table 3.2. Bootstrap support values (1000 replicates) for monophyly of clades I–V from maximum parsimony analyses of specific regions of MLO protein sequences.

MLO region analysed	Bootstrap support values				
	Clade I	Clade II	Clade III	Clade IV	Clade V
Full protein excl. N- and C- termini	100	99	92	100	100
Intra- and extracellular regions	100	100	78	100	100
Transmembrane regions	87	73	59	99	100
Intracellular regions	99	90	66	95	100
Extracellular regions	95	55	<50	99	100

Since monocots are believed to have diverged from dicots approximately 100-270 million years ago (Wolfe et al. 1989, Schneider-Poetsch et al. 1998), *Mlo*-like genes must have already existed in their common progenitor. In fact, it would appear that the age of this gene family is much older than the monocots and dicots. The monocot and dicot MLO sequences AtMLO4/ZmML04 and AtMLO1/ZmMLO8 group together as sister homologs with bootstrap values of 100 and 70 respectively (nodes A and B in Fig. 3.1). Unless these relationships are the result of horizontal gene transfer, the ages of these two nodes can be no younger than the 100-270 million year divergence time between monocots and dicots. Several ESTs have been identified for the gymnosperm *Pinus taeda* demonstrating presence of *Mlo* homologs in both subphyla of the spermatophyta (seed plants), angiosperms and gymnosperms, which are believed to have diverged from a common ancestor about 340-360 million years ago (Wolfe et al. 1989, Troitsky et al. 1991). Moreover, several ESTs (~20 out of ~65,000) with high sequence similarity to *Mlo* originate from the bryophyte *Physcomitrella patens*, and one (out of ~1,700 ESTs) from the moss *Ceratodon purpureus*. A maximum parsimony analysis of an alignment based on the regions corresponding to the partial *C. purpureus* sequence (68 amino acids of the C-terminus; Fig. 3.1) shows this sequence to fall within the diversity of monocots and dicots, with moderate bootstrap support for its placement within subfamily I. Bryophytes and tracheophytes (vascular plants) are believed to have diverged early in the evolution of green land plants between the mid-Ordovician and the early Silurian period, approximately 400-450 million years ago (Wolfe et al. 1989, Kenrick et al. 1997). Thus, unless this is the result of horizontal gene transfer, a common ancestor of both must already have possessed an *Mlo* homolog and the node uniting

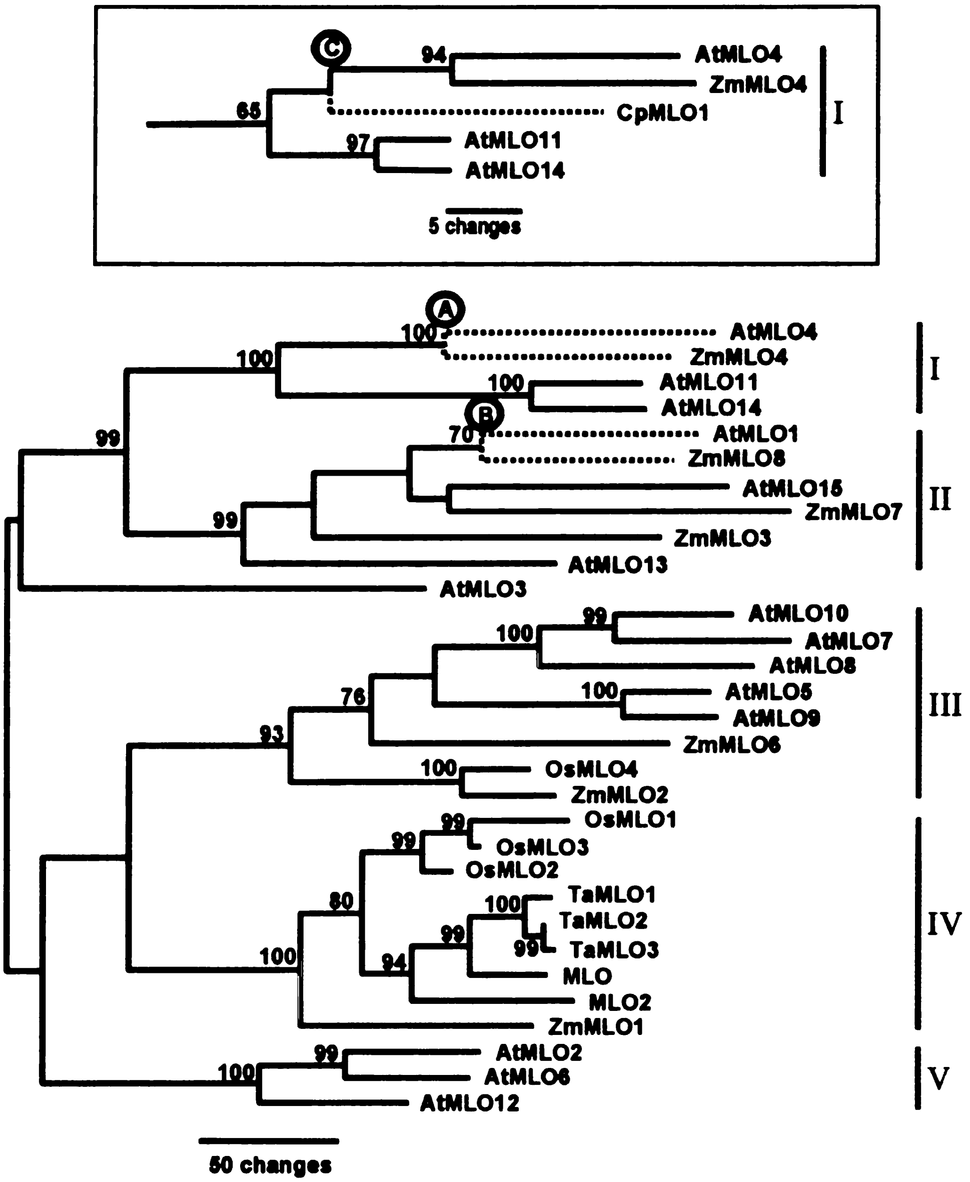

C
94
AtMLO4
ZmMLO4
65
CpMLO1
I
97
AtMLO11
AtMLO14
5 changes
A
100
AtMLO4
ZmMLO4
100
100
AtMLO11
AtMLO14
I
B
99
70
AtMLO1
ZmMLO8
AtMLO15
II
ZmMLO7
99
ZmMLO3
AtMLO13
AtMLO3
99
AtMLO10
100
AtMLO7
AtMLO8
100
AtMLO5
76
AtMLO9
III
93
ZmMLO6
100
OsMLO4
ZmMLO2
99
OsMLO1
99
OsMLO3
OsMLO2
80
100
TaMLO1
TaMLO2
IV
99
100
99
TaMLO3
94
MLO
MLO2
ZmMLO1
99
AtMLO2
100
AtMLO6
V
AtMLO12
50 changes

Fig. 3.1 Maximum parsimony phylogenetic analysis of amino acid sequence data for monocot and dicot MLO family members.

Maximum parsimony tree constructed from full-length amino acid sequence data for MLO genes, excluding N- and C-termini. Branch lengths are proportional to the amount of amino acid changes. Numbers at the nodes indicate bootstrap support values (1000 replicates) above 60. Roman numerals denote major clades (subfamilies) referred to in the text. Nodes A and B indicate monocot and dicot sister lineages (dashed lines) referred to in the text. The inset indicates the phylogenetic position (node C) of the bryophyte MLO sequence (CpMLO1) from a maximum parsimony analysis of an alignment of partial sequences corresponding to the 68 amino acids of CpMLO1. The analysis included all MLO sequences in the partial alignment but for clarity only clade I containing CpMLO1 is shown.

CpMLO1/AtMLO4/ZmMLO4 (node C in Fig. 3.1) can be no younger than the 400-450 million year divergence time between bryophytes and tracheophytes.

We conclude from this observation that the presence of *Mlo* genes can be traced back at least to the early evolutionary stages of land plant development. This implies an ancient and vital function for the MLO family in plants. EST database searches (http://www.kazusa.or.jp/en/plant) of the unicellular green alga *Chlamydomonas reinhardtii* (37,990 ESTs) and the marine red alga *Porphyra yezoensis* (10,154 ESTs) detected no *Mlo*-like sequences in these two species. This could be first evidence that *Mlo* emerged concurrently with the conquest of terrestrial habitats, although we cannot rule out the possibility that the number of currently available algal ESTs is too low to identify *Mlo*-like sequences.

Closely related members belonging to the same subfamily but originating from different species may be identified as orthologs with similar functions, as experimentally demonstrated for MLO, TaMLO2 and OsMLO2 (see above; Elliott et al. in press). Whether the observed clustering correlates generally with a common function of the members is currently under investigation.

A common scaffold topology accommodates two hypervariable domains

A hallmark of all MLO family members is the presence of seven TM domains. The predictions obtained for each of the full-size family members from Table 3.1 by using the TMHMM algorithm (Sonnhammer et al. 1998) exactly matched the 7TM topology determined experimentally for the barley MLO protein (Devoto et al. 1999). Similarly,

the predicted distribution of the amino acid residues with respect to the membrane is comparable to the barley protein: generally 50-60% of the protein is predicted to be cytoplasmic, 20-30% to be embedded in the membrane, and the rest is thought to be extracellular/lumenal. These observations indicate a shared scaffold topology for all MLO protein family members, consisting of seven TM helices, an N-terminal extracellular or lumenal end, three cytoplasmic and three extracellular/lumenal loops, and a cytoplasmic C-terminal tail (Fig. 3.2). Although a rice MLO homolog has also been shown to reside within the plasma membrane (Kim et al. 2002a), the scaffold topology does not provide conclusive evidence for a common subcellular localization. For simplicity, we refer in the following to "extracellular" rather than to "extracellular/lumenal" domains.

Another characteristic is the presence of four strictly conserved cysteine residues in extracellular loops 1 and 3 (Fig. 3.2). If these cysteine residues form (a) disulfide bridge(s) either with each other or with the two other invariant extracellular cysteines, this domain could subsequently form an exposed loop/ligand binding site. This is frequently found in mammalian 7TM receptors to stabilize the relative arrangement of the TM helices to each other (Probst et al. 1992, Strader et al. 1994). Extraordinary length variability occurs between cysteine residues 99 and 115 in extracellular loop 1, contributing to an exceptional sequence variation in this region among family members (Fig. 3.3A). The C-terminus defines the second domain that is highly variable both in sequence and length (ranging from 55 to 253 amino acid residues, Fig. 3.3B). However, the first ~25 residues proximal to TM VII are rather conserved, harboring the recently discovered calmodulin binding site present in MLO proteins (Fig. 3.2B; Kim et al. 2002a

and b). A hallmark of this binding site is a strictly conserved tryptophan residue that has been demonstrated to be essential for the interaction with calmodulin (Fig. 3.2 and 3.3B; Kim et al. 2002a and b).

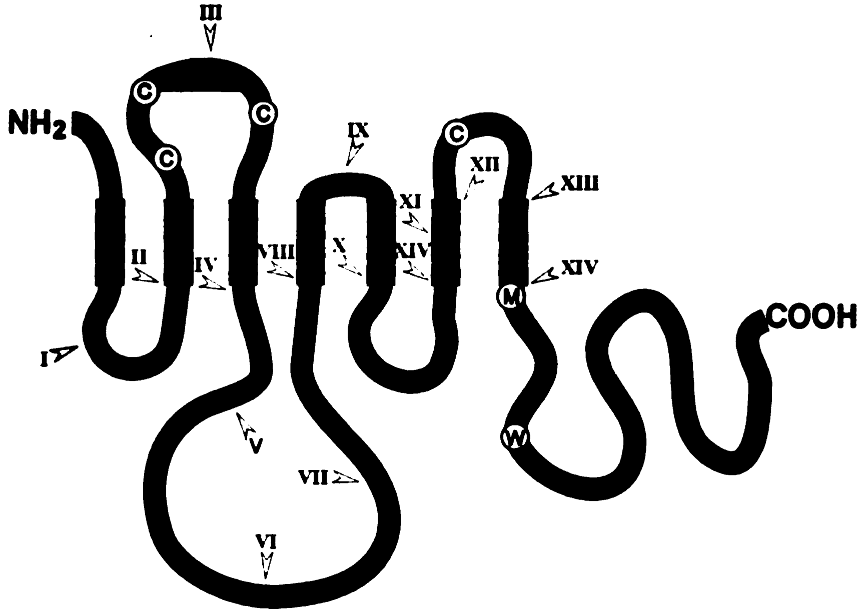


Fig. 3.2. Scheme of the MLO protein.

Grey boxes designate the seven TM helices. Arrows indicate the position of splice junctions (exon/exon junctions at protein level), with the corresponding introns numbered by roman numerals. C, M and W denote conserved cysteine, methionine and tryptophane residues, respectively.

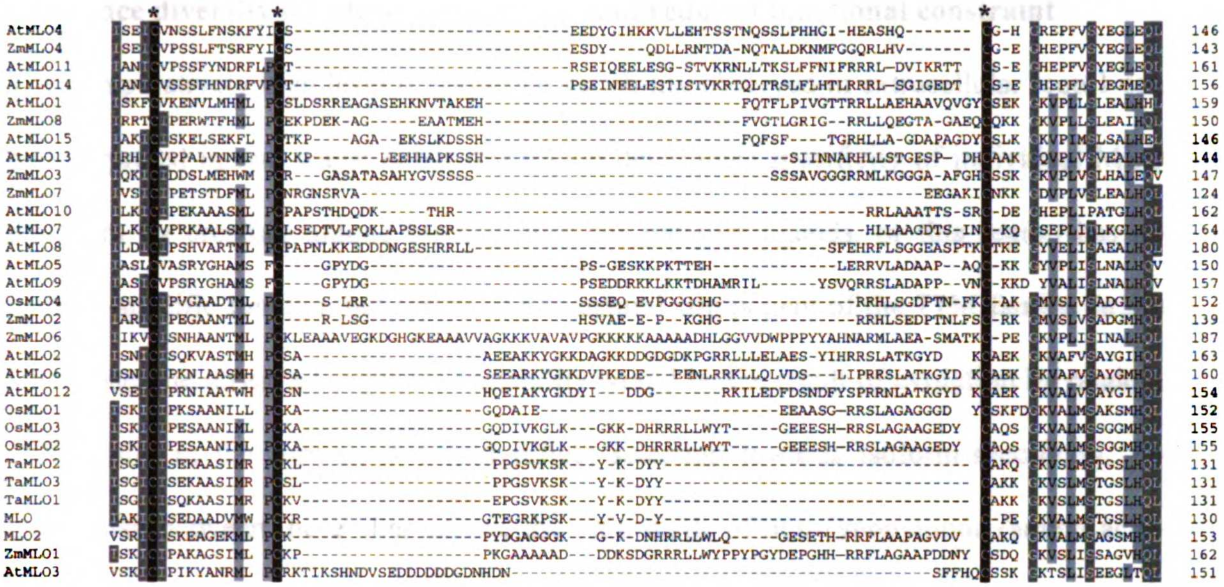


Fig. 3.3. Multiple sequence alignment of MLO proteins.

A. Alignment of amino acid sequences corresponding to the first extracellular loop.

B. Alignment of amino acid sequences corresponding to TM VII and the C-terminal tail.

Sequences were aligned using PileUp (Winsconsin Package Version 10.0, Genetics computer group, Madison, WI, USA); spaces were manually introduced to increase the similarity in the alignment. Shading indicates degree of conservation between amino acids. Identical amino acid residues (100% conserved) are shaded in black; 80% or greater conserved, 60% or greater conserved, and less than 60% conserved are indicated by dark gray, light gray, and white, respectively.

Numbers indicate amino acid positions within the protein; * indicate conserved cysteine residues; black and grey triangles indicate the methionine corresponding to the start of the last exon and the conserved tryptophane residue of the calmodulin binding domain, respectively.

Sequence diversity in extracellular loop 1 and reduced functional constraint

The comparatively high level of sequence variability observed in extracellular loop 1 can be interpreted in two ways: either this region determines specificity of individual MLO members by creating unique binding sites for putative ligands, or this region has no isoform-specific function but serves as a structural component of the 7TM family. In the latter case, the observed sequence variability would be the result of evolution by random drift, while in the former, it would reflect selection towards isoform-specificity. To distinguish between these alternatives, the ratio d_N/d_S of non-synonymous (amino acid-changing; d_N) and synonymous (silent; d_S) substitutions per non-synonymous and synonymous sites is a suitable indicator. Pseudogenes without any evolutionary selective pressure will accumulate neutral and amino acid-changing substitutions in their DNA sequence with the same frequency, resulting in a d_N/d_S ratio of approximately one. In contrast, in the majority of genes most of the occurring non-synonymous changes are probably deleterious, resulting in purifying counter-selection. In these cases, synonymous substitutions take place more often than non-synonymous ones, resulting in a d_N/d_S ratio below one. As a third possibility, certain coding regions are selected for extraordinary high rates of non-synonymous substitutions (resulting in a d_N/d_S ratio >1). This behavior is true for fast evolving genes that underlie adaptive molecular evolution as for example several surface antigens of pathogens and the matching defense systems in the corresponding hosts (Yang and Bielawski 2000). Since this method provides reliable results only if the sequences investigated are neither too similar nor too different (Yang and Bielawski 2000), we first had to select suitable sequences. Known full-size MLO sequences were unsuitable because they are highly divergent in extracellular loop 1

(Fig. 3.3A). We PCR-amplified a fragment of the *Mlo* genomic sequence (corresponding to extracellular loop 1 and some flanking amino acid residues) from eight different species of the genus *Hordeum* (Materials and Methods, Table 3.3 and Fig. 3.4). In two cases, we obtained two distinct sequences each, likely reflecting the polyploid nature of these species. The resulting predicted amino acid sequences are only moderately divergent in extracellular loop 1 and thus ideally suited for d_N/d_S analyses (compare Fig. 3.3A and 3.4).

Table 3.3. Sequences of *Hordeum* species used for the d_N/d_S analysis

Species	**Ploidy**[a]	**GenBank accession no.**
H. vulgare	diploid	Z83834
H. vulgare f. *agriocrithon*	diploid	AY090646
H. vulgare ssp. *spontaneum*	diploid	AY090647
H. brevisbulatum	diploid, tetraploid and hexaploid	AY090638, AY090639
H. bulbosum	diploid and tetraploid	AY090641, AY090642
H. chilense	diploid	AY090643
H. jubatum	tetraploid	AY090640
H. murinum ssp. *murinum*	tetraploid	AY090645
H. murinum ssp. *leporinum*	tetraploid and hexaploid	AY090644

[a] according to von Bothmer et al. 1995

We calculated d_N/d_S ratios for each possible pair of ten amplified and two previously known sequences of (i) the region corresponding to amino acid residues 69-145 of barley MLO, (ii) extracellular loop 1 excluding the region between conserved cysteines 86 and 114, and (iii) the region between conserved cysteines 86 and 114. The average d_N/d_S ratio values for each of these are 0.138, $\sigma = 0.048$ (i), 0.154, $\sigma = 0.054$ (ii), and 0.275, $\sigma = 0.170$ (iii). All of these values are well below 1 indicating functional constraint on the evolution of the DNA sequences and purifying selection. However, it appears that functional constraint is less for the region between conserved cysteine residues 86 and 114 in extracellular loop 1 as the average d_N/d_S ratio for this section is almost two times higher than that of its 5' and 3' flanking sequences (0.275 versus 0.154, respectively), although this difference is not statistically significant. This result can be interpreted in two ways. It may indicate that relaxed constraint in DNA evolution causes over long periods of time the sequence variation found among compiled MLO family members. Alternatively, in this particular case, the d_N/d_S ratio might not be a reliable indicator for the molecular mechanism leading to the observed variability. It will be interesting to find out whether differences in this region correspond to the ability to bind diverse interacting partners in the extracellular space.

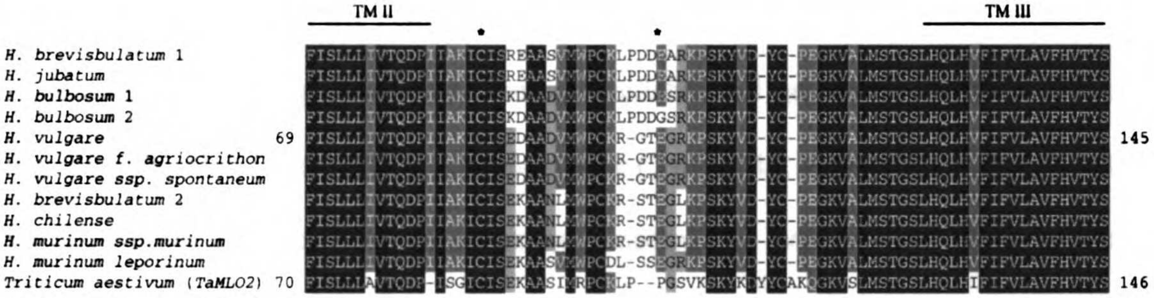

TM II
TM III
H. brevisbulatum 1 FISLLLIVTQDPIIAKICISREAASVMWPCKLPDDEARKPSKYVD-YC-PEGKVALMSTGSLHQLHVFIFVLAVFHVTYS
H. jubatum FISLLLIVTQDPIIAKICISREAASVMWPCKLPDDEARKPSKYVD-YC-PEGKVALMSTGSLHQLHVFIFVLAVFHVTYS
H. bulbosum 1 FISLLLIVTQDPIIAKICISKDAADVMWPCKLPDDESRKPSKYVD-YC-PEGKVALMSTGSLHQLHVFIFVLAVFHVTYS
H. bulbosum 2 FISLLLIVTQDPIIAKICISKDAADVMWPCKLPDDGSRKPSKYVD-YC-PEGKVALMSTGSLHQLHVFIFVLAVFHVTYS
H. vulgare 69 FISLLLIVTQDPIIAKICISEDAADVMWPCKR-GTEGRKPSKYVD-YC-PEGKVALMSTGSLHQLHVFIFVLAVFHVTYS 145
H. vulgare f. agriocrithon FISLLLIVTQDPIIAKICISEDAADVMWPCKR-GTEGRKPSKYVD-YC-PEGKVALMSTGSLHQLHVFIFVLAVFHVTYS
H. vulgare ssp. spontaneum FISLLLIVTQDPIIAKICISEDAADVMWPCKR-GTEGRKPSKYVD-YC-PEGKVALMSTGSLHQLHVFIFVLAVFHVTYS
H. brevisbulatum 2 FISLLLIVTQDPIIAKICISEKAANLMWPCKR-STEGLKPSKYVD-YC-PEGKVALMSTGSLHQLHVFIFVLAVFHVTYS
H. chilense FISLLLIVTQDPIIAKICISEKAANLMWPCKR-STEGLKPSKYVD-YC-PEGKVALMSTGSLHQLHVFIFVLAVFHVTYS
H. murinum ssp.murinum FISLLLIVTQDPIIAKICISEKAANLMWPCKR-STEGLKPSKYVD-YC-PEGKVALMSTGSLHQLHVFIFVLAVFHVTYS
H. murinum leporinum FISLLLIVTQDPIIAKICISEKAASVMWPCDL-SSEGRKPSKYVD-YC-PEGKVALMSTGSLHQLHVFIFVLAVFHVTYS
Triticum aestivum (TaMLO2) 70 FISLLLAVTQDP-ISGICISEKAASIMRPCKLP--PGSVKSKYKDYYCAKQGKVSLMSTGSLHQLHIFIFVLAVFHVTYS 146

Fig. 3.4. Amino acid sequence alignment of MLO sequences used for the d_N/d_Sanalysis

Amino acid sequences corresponding to extracellular loop 1 and flanking regions from 11 presumptive orthologs of nine different species of the genus *Hordeum* and a wheat sequence were aligned using ClustalW. Identical amino acid residues (100% conserved) are shaded in black; 80% or greater conserved, 60% or greater conserved, and less than 60% conserved are indicated by dark gray, light gray, and white, respectively.. The two asterisks indicate conserved cysteines that are at position 86 and 114 in barley MLO.

Structural organization of Mlo genomic sequences provides further evidence for a monophyletic origin of the gene family

A comparison of the gene structure among available full genomic sequences of family members revealed 11 to 14 introns per *Mlo* gene (Table 3.1). Most of the introns are 80 to 90 nucleotides in size, with no sequence conservation even among phylogenetically closely related members. It is noticeable that in all but one case the exon/exon junctions map exactly at the identical position at the corresponding protein level, supporting a monophyletic origin for the gene family (Fig. 3.2). The only exception is represented by intron V that is located at a slightly different position in *AtMlo1*, *13*, and *15*. Intron VI is absent in *AtMlo2* and *AtMlo6*, while intron XI is missing in *AtMlo1*, *13*, and *15*. These observations are in full agreement with the phylogenetic analysis (see above and Fig. 3.1) suggesting that highly similar members within Arabidopsis did not arise by convergence from different progenitor sequences but diverged from a single common ancestor gene. The C-terminal tails are always encoded by a single exon, invariably starting with a consensus translational initiation sequence including the start codon 'ATG' (Fig. 3.2 and 3.3B). Whether this reflects an ancient gene shuffling event remains speculative.

The splice junctions in the gene family mainly map to the boundaries between the encoded loop and transmembrane regions (Fig. 3.2). Eight of the 14 exon/exon junctions are located proximal or distal to the transmembrane helical termini. Only one TM helix (VI) is interrupted by a splice junction. The remaining junctions are located within extracellular loop 1, intra- and extracellular loop 2, and TM helix VI. No exon-exon junction was observed in the amino- and the carboxyl-terminal ends of the family

members proximal to the first TM helix or distal to TM VII. The fact that individual TM helices are encoded by single exons is common to other polytopic membrane proteins (Argos and Rao 1985, Miao and Verma 1993). This is thought to reflect their role as an evolutionary unit that is subject to severe selection constraint to maintain the structurally stable, multihelical transmembrane core. Such a unit may serve as module to create variability in the number of TM helices of polytopic membrane proteins (e.g. by exon shuffling).

AtMlo distribution in the Arabidopsis genome

It has been demonstrated recently that most of the genome of *Arabidopsis thaliana* is internally duplicated, indicating Arabidopsis as a potential ancient tetraploid species (Blanc et al. 2000, The Arabidopsis Genome Initiative 2000, Vision et al. 2000). Additionally, Vision et al. (2000) provided evidence that the current state of the Arabidopsis genome may result from at least four different large-scale duplication events that took place 100 to 200 million years ago. These duplication processes must have also involved chromosome fusions resulting in extended genomic regions in which number, order, and orientation of duplicated genes are preserved. After duplication, affected regions were subject to extensive subchromosomal rearrangements, such as inversions, translocations and loss or transposition of single genes or groups of genes.

We investigated the distribution of *AtMlo* genes in extended duplicated genomic regions in order to identify putative functionally redundant copies of *AtMlo* genes. For this analysis we used the template map of Arabidopsis genomic duplications described in

Blanc *et al.* (2000) because start and end of the copied regions are exactly designated by particular BAC clones. We found that *Mlo* genes are located on all five chromosomes without any obvious clustering. With two exceptions (*AtMlo9* and *13*), all *AtMlo* genes are located within regions that are supposed to have undergone a previous large-scale duplication event (Fig. 3.5). Unexpectedly, *AtMlo* genes were always found as a single copy in the duplicated areas, except *AtMlo2* and *AtMlo6* for which number, order, and orientation of flanking genes are rather conserved. Although it is known that less than half of the genes (37-47%, depending on significance criteria) in the duplicated areas are conserved in their corresponding copy region (Bancroft, 2001), *AtMlo* genes behave differently because of only a single recognizable duplication. Whether this indicates constraints in copy numbers or exceptionally high micro-translocation/deletion events cannot be resolved. Taken together, this approach identifies only AtMLO2 and AtMLO6 as the result of a large-scale duplication event. It should be interesting to find out whether these two genes are functionally redundant or whether the few sequence differences lead to functional diversification.

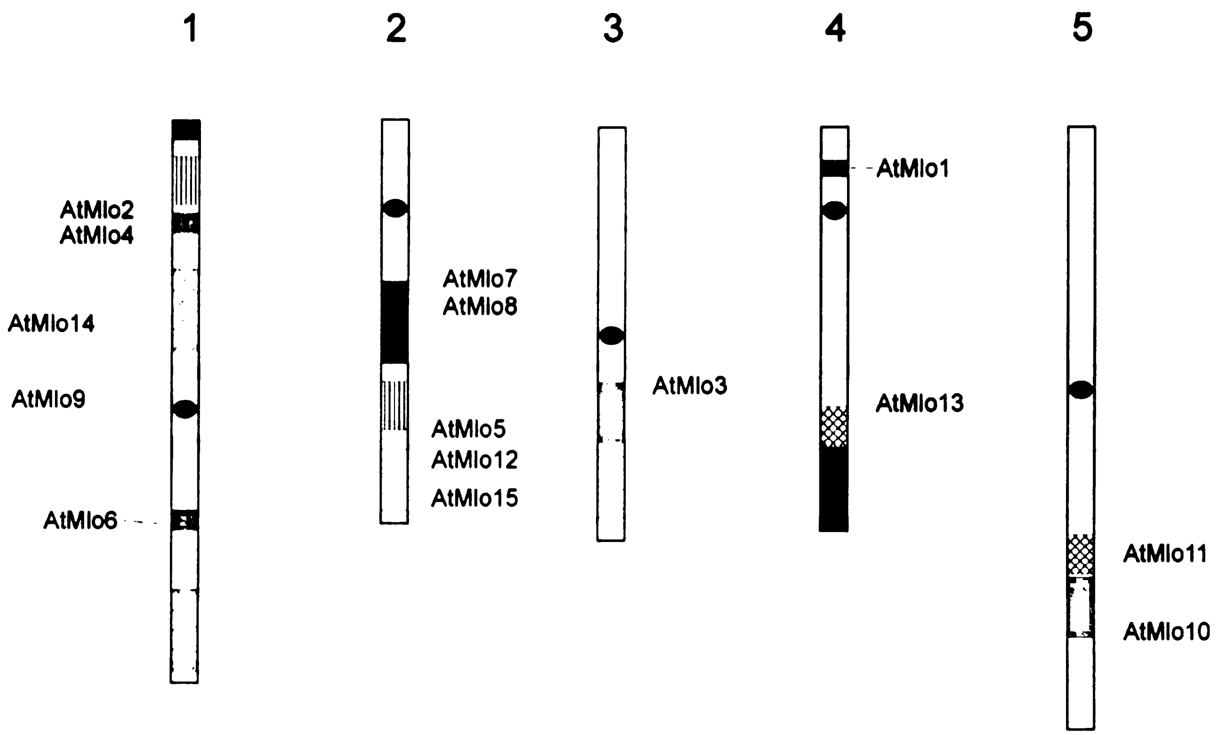


Fig. 3.5. Distribution of *AtMlo* members in the Arabidopsis genome.

The five chromosomes of *Arabidopsis thaliana* are schematically represented by rectangles numbered from 1-5. Centromeric regions are indicated by black ovals. Marked blocks indicate areas of large-scale genome duplications. Relative positions of the 15 *AtMlo* genes are shown. (Adapted from Blanc et al. 2000).

Co-evolution among domains of MLO proteins

Recently, Goh et al. (2000) have developed an algorithm that allows the identification of protein-protein interaction pairs and can be adapted to the assessment of intramolecular co-evolution of peptide domains within a single protein family. The method is based on the assumption that if there are two domains within a single protein that have to act co-operatively for proper function, evolutionary changes of the amino acid sequence within one of the domains will either result in counter-selection or in compensating changes in the amino acid sequence of the other domain. In terms of evolution, these two domains will evolve in a coordinated manner resulting in a linked phylogenetic relationship. If there is no co-operation between the two domains, they are believed to evolve independently resulting in an unlinked phylogenetic relationship. The algorithm has been used by Pazos and Valencia (2001) to test the impact of the method by analyzing potential intramolecular interactions of structural domains in bipartite proteins and by investigating known protein-protein interaction pairs. The authors conclude from their results that the procedure is capable to detect true interactions in >66% of the cases if a correlation >0.8 is detected.

We dissected 31 full-length sequences of MLO proteins into their single peptide domains. This procedure resulted in 15 sets of peptide sequences, representing the N- and C-termini, the seven TM regions, the three cytoplasmic, and the three extracellular loops. We paired each set of peptide sequences with each other and calculated correlation coefficients for all 105 possible pairings (Fig. 3.6 and Materials and Methods). The

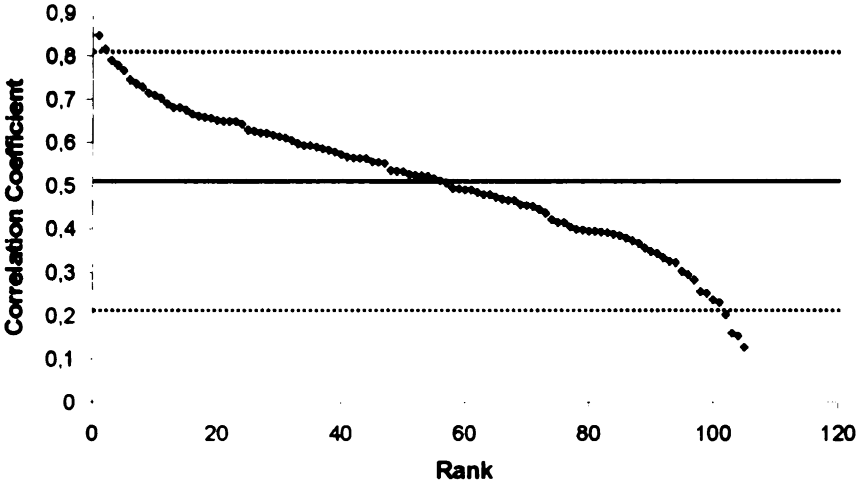


Fig. 3.6. Inter-domain correlation analysis of MLO.

Correlation coefficients of all 105 inter-domain pairings of the 15 sets of peptide domains from 31 MLO proteins were plotted against the relative ranking (ranging from 1 to 105) of the respective pair. Mean value and 1.96x standard deviations are indicated by a bold line and dotted lines, respectively.

In Table 3.4 we have listed the top five pairings with the highest correlation coefficients that we will discuss in detail. All of them have values close to or even above the 1.96 times standard deviation boundary (marking significant values with a probability value of p<0.05). This indicates that co-evolution between the respective peptide domains is likely. Among these top five pairs, the three possible combinations between the cytoplasmic domains IC2, IC3, and the C-terminus have the highest scores (Table 3.4) of about 0.8, a value that has been suggested to be a good empirical cut-off to indicate with a high probability true positive interactions (Pazos and Valencia, 2001). The following two pairs both indicate also a possible co-evolution of IC1 with loops IC2 and IC3 (Table 3.4). Taken together, the analysis provides evidence for co-evolution of all cytoplasmic loops with the C-terminus, showing a particular emphasis on IC2, IC3 and the C-terminus. Probable co-evolution between the cytoplasmic domains of MLO suggests interplay of these domains and interaction with putative partner(s) for MLO protein function. Although other scenarios are possible, the most likely interpretation is related to a c onserved interaction of the c ytoplasmic domains with a common binding partner. An analogous situation has been demonstrated experimentally for the well-characterized family of GPCRs in binding heterotrimeric G-proteins (reviewed in Hamm 1998). The relative absence of correlations joining the extracellular domains could relate to the heterogeneity of presumptive ligands that might bind and activate MLO proteins.

Table 3.4. Correlation coefficients of the co-evolution analysis of MLO protein domains

Rank	Pair	Correlation coefficient
1	IC3/C-terminus	0.85
2	IC2/IC3	0.82
3	IC2/C-terminus	0.79
4	IC1/IC2	0.78
5	IC1/IC3	0.77

IC; intracellular loop

Gen Bank accession numbers

GenBank accession numbers for newly deposited sequences are as follows: Z95353 (*OsMlo1*), AF384030 (*OsMlo2*), AF361933 (*TaMlo1*), AF361932 (*TaMlo2*), Z95352 and AF369563-AF369576 (*AtMlo1-AtMlo15*), AY029312-AY029320 (*ZmMlo1-ZmMlo9*), and AY090638-AY090647 (*Hordeum* species of Table 3)

Acknowledgements

We thank our colleague Günter Theissen for helpful and critical comments on the manuscript. We are grateful to Brigitte Schauf (Max-Planck Institut für Züchtungsforschung), Robin Wineland (Pioneer) and Wayne Powell (DuPont) for their assistance.

This work was supported by grants from the Gatsby Charitable Foundation, the BBSRC and the Max-Planck Society to P. S.-L. A. D. was supported by an EU fellowship (Training and Mobility of Researcher Program in Biotechnology), H.A. H. by an EMBO fellowship. C.-S. G. and F.E. C. were supported by FEC GM39900 from the National Institute of Health.

References

Argos P, Rao JKM (1985) Relationships between exons and the predicted structure of membrane-bound proteins. Biochim Biophys Acta 827: 283-297

Bancroft I (2001) Duplicate and diverge: the evolution of plant genome microstructure. Trends Genet 17: 89-93

Blanc G, Barakat A, Guyot R, Cooke R, Delseny I (2000) Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell 12: 1093-1101

Bockaert J, Pin JP (1999) Molecular tinkering of G protein-coupled receptors: an evolutionary success. EMBO J 18: 1723-1729

Büschges R, Hollricher K, Panstruga R, Simons G, Wolter M, Frijters A, van Daelen R, van der Lee T, Diergaarde P, Groenendijk J, Töpsch S, Vos P, Salamini F, Schulze-Lefert P (1997) The barley *Mlo* gene: a novel control element of plant pathogen resistance. Cell 88: 695-705

Devoto A, Piffanelli P, Nilsson I, Wallin E, Panstruga R, von Heijne G, Schulze-Lefert P (1999) Topology, subcellular localization, and sequence diversity of the *Mlo* family in plants. J Biol Chem 274: 34993-35004

Elliott C, Zhou F, Spielmeyer W, Panstruga R, Schulze-Lefert P (in press) Functional conservation of wheat and rice Mlo orthologs in defence modulation to the powdery mildew fungus. Mol Plant-Microbe Interact (in press)

Freialdenhoven A, Peterhänsel C, Kurth J, Kreuzaler F, Schulze-Lefert P (1996) Identification of genes required for the function of non-race- specific *mlo* resistance to powdery mildew in barley. Plant Cell 8: 5-14

Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE (2000) Co-evolution of proteins with their interaction partners. J Mol Biol 299: 283-293

Hamm H E (1998) The many faces of G protein signaling. J Biol Chem 273: 669-672

Josefsson LG, Rask L (1997) Cloning of a putative G-protein-coupled receptor from *Arabidopsis thaliana*. Eur J Biochem 249: 415-420

Kenrick P, Crane PR (1997) The origin and early evolution of plants on land. Nature 389: 33-39

Kim MC, Lee SH, Kim JK, Chun HJ, Ok HM, Moon BC, Kang CH, Chung WS, Park CY, Choi MS, Kang YH, Koo SC, KooYC, Jung JC, Schulze-Lefert P, Cho MJ (2002a) *Mlo*, a modulator of plant defense and cell death, is a novel calmodulin-binding protein: isolation and characterization of a rice *Mlo* homologue. J Biol Chem 277: 19304-19314

Kim MC, Panstruga R, Elliott C, Müller J, Devoto A, Yoon HW, Park HC, Cho MJ, Schulze-Lefert P (2002b) Calmodulin interacts with MLO protein to regulate defence against mildew in barley. Nature 416: 447-450

Miao GH, Verma DPS (1993) Soybean nodulin-26 gene encoding a channel protein is expressed only in the infected cells of nodules and is regulated differently in roots of homologous and heterologous plants. Plant Cell 5: 781-794

Pazos F, Valencia A (2001) Similarity of Phylogenetic trees as indicator of protein-protein interaction. Prot Eng. 14: 609-614

Peterhänsel C, Freialdenhoven A, Kurth J, Kolsch R, Schulze-Lefert P (1997) Interaction analyses of genes required for resistance responses to powdery mildew in barley reveal distinct pathways leading to leaf cell death. Plant Cell 9: 1397-1409

Piffanelli P, Zhou F, Casais C, Orme J, Schaffrath U, Collins N, Panstruga R, Schulze-Lefert P (in press) The barley MLO modulator of defence and cell death is responsive to biotic and abiotic stress stimuli. Plant Physiol. (in press)

Plakidou-Dymock S, Dymock D, Hooley R (1998) A higher plant seven-transmembrane receptor that influences sensitivity to cytokinins. Curr Biol 8: 315-324

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1998) Numerical recipes in C. Cambridge University Press, Cambridge, UK

Probst WC, Snyder LA, Schuser DI, Brosius J, Sealfon SC (1992) Sequence alignment of the G-protein coupled receptor superfamily. DNA Cell Biol 11: 1-20

Schneider-Poetsch HAW, Kolukisaoglu U, Clapham DH, Hughes J, Lamparter T (1998) Non-angiosperm phytochromes and the evolution of vascular plants. Physiol Plant 102: 612-622

Shirasu K, Nielsen K, Piffanelli P, Oliver R, Schulze-Lefert P (1999) Cell-autonomous complementation of *mlo* resistance using a biolistic transient expression system. Plant J 17: 293-299

Sonnhammer ELL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. In: Glasgow J, Lathorp R, Littlejohn T, Major F, (eds) Proc Sixth Int Conf on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, pp.175-182

Strader CD, Fong TM, Tota MR, Underwood D, Dixon RAF (1994) Structure and function of G-protein-coupled receptors. Annu Rev Biochem 63: 101-132

Swofford DL (1998) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer, Sunderland, MA

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796-815

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 22: 4673-4680

Troitsky AV, Melekhovets YF, Rakhimova GM, Bobrova VK, Valiejoroman KM, Antonov AS (1991) Angiosperm origin and early stages of seed plant evolution deduced from rRNA sequence comparisons. J Mol Evol 32: 253-261

Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in Arabidopsis. Science 290: 2114-2117

von Bothmer R, Jacobsen N, Baden C, Jørgensen RB, Linde-Laursen I (1995) An ecogegraphical study of the genus *Hordeum*. Systematic and ecogeographic studies on crop genepools 7. International Plant Genetic Resources Institute, Rome (2nd edition)

Wolfe KH, Gouy ML, Yang YW, Sharp PM, Li WH (1989) Date of the monocot dicot divergence estimated from chloroplast DNA-sequence data. Proc Natl Acad Sci 86: 6201-6205

Wolter M, Hollricher K Salamini F, Schulze-Lefert P (1993) The *Mlo* resistance alleles to powdery mildew infection in barley trigger a developmentally controlled defense mimic phenotype Mol Gen Genet 239: 122-128.

Yang ZH. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13: 555–556.

Yang ZH, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496-503

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17: 32-43.

Chapter 4

Co-Evolutionary Analysis of Interacting Proteins Reveals Insights into into Protein-Protein Interactions

This chapter is in press as:

Goh CS, Cohen FE. Co-Evolutionary Analysis of Interacting Proteins Reveals Insights into Protein-Protein Interactions. *J Mol Biol.*

ABSTRACT

Protein-protein interactions play crucial roles in biological processes. Experimental methods have been developed to survey the proteome for interacting partners and some computational approaches have been developed to extend the impact of these experimental methods. Computational methods are routinely applied to newly discovered genes to infer protein function and plausible protein-protein interactions. Here we develop and extend a quantitative method that identifies interacting proteins based upon the correlated behavior of the evolutionary histories of protein ligands and their receptors. We have studied six families of ligand-receptor pairs including: the syntaxin/unc-18 family, the GPCR/G-α's, the TGF-β/TGF-β receptor system, the immunity/colicin domain collection from bacteria, the chemokine/chemokine receptors, and the VEGF/VEGF receptor family. For correlation scores above a defined threshold, we were able to find an average of 79% of all known binding partners. We then applied this method to find plausible binding partners for proteins with uncharacterized binding specificities in the syntaxin/Unc-18 protein and TGF-β/TGF-β receptor families. Analysis of the results show that co-evolutionary analysis of interacting protein families can reduce the search space for identifying binding partners by not only finding binding partners for uncharacterized proteins but also recognizing potentially new binding partners for previously characterized proteins. We believe that correlated evolutionary histories provide a route to exploit the wealth of whole genome sequences and recent systematic proteomic results to extend the impact of

these studies and focus experimental efforts to categorize physiologically or pathologically relevant protein-protein interactions.

***Keywords*: co-evolution; protein-protein interaction; TGF-β; syntaxin; chemokines; G-protein coupled receptors; colicin; VEGF**

INTRODUCTION

Identification of protein-protein interactions is essential for the understanding of various cellular processes including systems involved in metabolic, signaling, and regulatory pathways. Most of our understanding of these interactions comes from high-throughput two-hybrid studies, mass spectrometry, or traditional biochemical and genetic approaches.[1-5] In an effort to complement experimental methods, several computational approaches have been proposed to predict protein interactions by incorporating information found in both families of sequences and whole genomes.[6-10] While these methods can be useful in defining functions of a certain subset of uncharacterized genes in completed genomes, we believe that the co-evolution of ligand-receptor pairs in evolving organisms provide a powerful and orthogonal approach to identifying protein-protein interactions.

Previously, we reported a method for quantitating the co-evolution between a family of protein ligands and their receptors to identify the special pairing of interacting ligands and receptors.[11] Pazos & Valencia (2001)[12] tested this hypothesis on a large collection of protein systems found in the *E.coli* genome and were able to demonstrate the utility of this method. Due to the large genomic data set and necessity for automation, the authors were only able to incorporate orthologous information (one homologous protein per species). Hence, while their approach was able to identify protein families that could interact, it could not recognize specific interacting partners . between the two protein families. We had previously suggested that potential binding partners could be inferred through the visual inspection of the phylogenetic trees.[11]

Despite the overall effectiveness of this approach to infer binding partners, the method was dependent to some extent on the phylogenetic method used as well as the subjective nature of visual inspection. Therefore we have developed a method that quantitatively infers binding proteins using the correlation between sequence similarity distance matrices constructed for specific protein families. This algorithm provides a more accurate measure of proteins that co-evolve in order to maintain their interactions.

Our approach is based upon the pattern of evolutionary distances between a particular protein and its family members quantified by the sequence identity scoring function in ClustalW.[13] If co-evolution is relevant, a ligand-receptor pair should occupy related positions in their phylogenetic trees. Previous results[11,12] have shown that for ligand-receptor pairs that are part of most large protein families, the correlation between their phylogenetic distance matrices is significantly greater than for unrelated protein families. Here we show that within these correlated phylogenetic trees, the protein pairs that bind have a higher correlation between their phylogenetic distance matrices than other homologs drawn from the ligand and receptor families that do not bind. By calculating the correlation of each protein pair and by incorporating experimentally determined binding data, we can not only quantitatively infer interacting partners for orphan ligands or receptors (proteins with uncharacterized binding specificity) but also identify additional binding partners for characterized proteins. We tested this hypothesis on six protein-protein interaction systems: the syntaxin/unc-18 protein families, the adrenergic receptors and their G-α subunits, the TGF-β proteins and their receptors, the colicin/immunity protein families, the chemokines and their receptors, and the VEGF proteins and their receptors. Each of these systems illustrates that proteins can co-evolve

in order to maintain their binding interfaces and their functional role in cellular physiology.

RESULTS

Syntaxin and Unc-18 Protein Families

The syntaxin family belongs to the t-SNARE subfamily of the SNARE superfamily and is involved in mediating vesicle trafficking.[14-17] The syntaxins have also been shown to form complexes with proteins of the Sec1 (mUNC-18) family.[18,19] Sec1 proteins are cytosolic proteins that play an essential role in vesicle trafficking. They are believed to act as chaperones that put syntaxins into conformations that are required to form a SNARE complex with other SNAREs.[15,20,21] Identifying the interactions between the syntaxin family and the unc-18 family can aid in understanding these complex molecular assemblies. The syntaxin and unc-18 protein families are an example of the co-evolution of an interacting protein-protein system.

Inferred Binding Partners for Selected Proteins

In this analysis, we used PSI-BLAST to automatically gather 37 distinct proteins from the syntaxin family. By including species variants, a total of 86 sequences were culled from the database. Of the 37 proteins in the syntaxin family, 15 of these proteins have known binding partners. Correspondingly, 54 sequences of the unc-18 family consisting of 8 proteins with characterized binding specficity and 3 uncharacterized proteins were retrieved. The unc-18 protein family was selected as the query family because it

contained a smaller number of proteins than the syntaxin family. To understand the approximate range of correlation scores, each of the 8 characterized proteins with known binding information in the unc-18 family was used as a query protein. For each query, we removed all known binding information for that protein and used the correlated evolution method to infer its binding partners (Table 4.1).

Table 4.1. Syntaxin-Unc18 Protein Family Binding Pairs

Unc-18 Proteins	**Syntaxins**
Unc-18	**Syntaxin 1a, Syntaxin 1b, Syntaxin 1c, Syntaxin 2, Syntaxin 3**
Unc-18b	**Syntaxin 3**
Unc-18c	**Syntaxin 2, Syntaxin 4**
KEULE	**Knolle**
Sly1	**Syntaxin 5, Syntaxin 17, Sed5p**
VPS45	**Pep12p, Tlg2p, Syntaxin 6**
VPS33	**Vam3p, Syntaxin 7**

These experimentally determined binding partners[15,64-74] were used to calculate the correlation coefficient and to infer additional binding pairs (see Methods).

The overall correlation value measuring the co-evolution of binding specificity for the unc-18 family and the syntaxin family was 0.82. This value was calculated as previously described in Goh *et al.*[11] The averaged correlation values of the true and unknown pairs were calculated for each protein query (Figure 4.1a). In an effort to avoid undiscovered but evolutionarily expected pairings, the correlations for the unknown pairs only included pairs with characterized proteins and did not contain pairs with orphans. For all the query proteins in the unc-18 family, the known binding pairs had an average correlation above 0.9 and the presumably false (unknown) partners had an average correlation value below 0.7. These results reflect that Unc-18b and Unc-18c have higher average correlation values for their unknown pairs, most likely because they are polyfunctional and share several binding partners with Unc-18.

Clearly, an objective quantitative measure of binding specificity can provide a faster and more precise measure than the visual inspection of the phylogenetic trees. One example is the query for the binding partner(s) of the KEULE protein in the unc-18 family (Figure 4.2). By visual inspection of the syntaxin tree (Figure 4.2a), the partner(s) for the KEULE protein could be any of the 10 proteins ranging from the accession number 11358872 protein to accession number 7447078 protein. In Figure 4.3, our quantitative results indicate that Knolle, the cognate partner o f KEULE, and its orthologue a re the second and third on the list. The most likely KEULE-binding partner is an uncharacterized protein, a Syr1-like protein (GI #4262161).

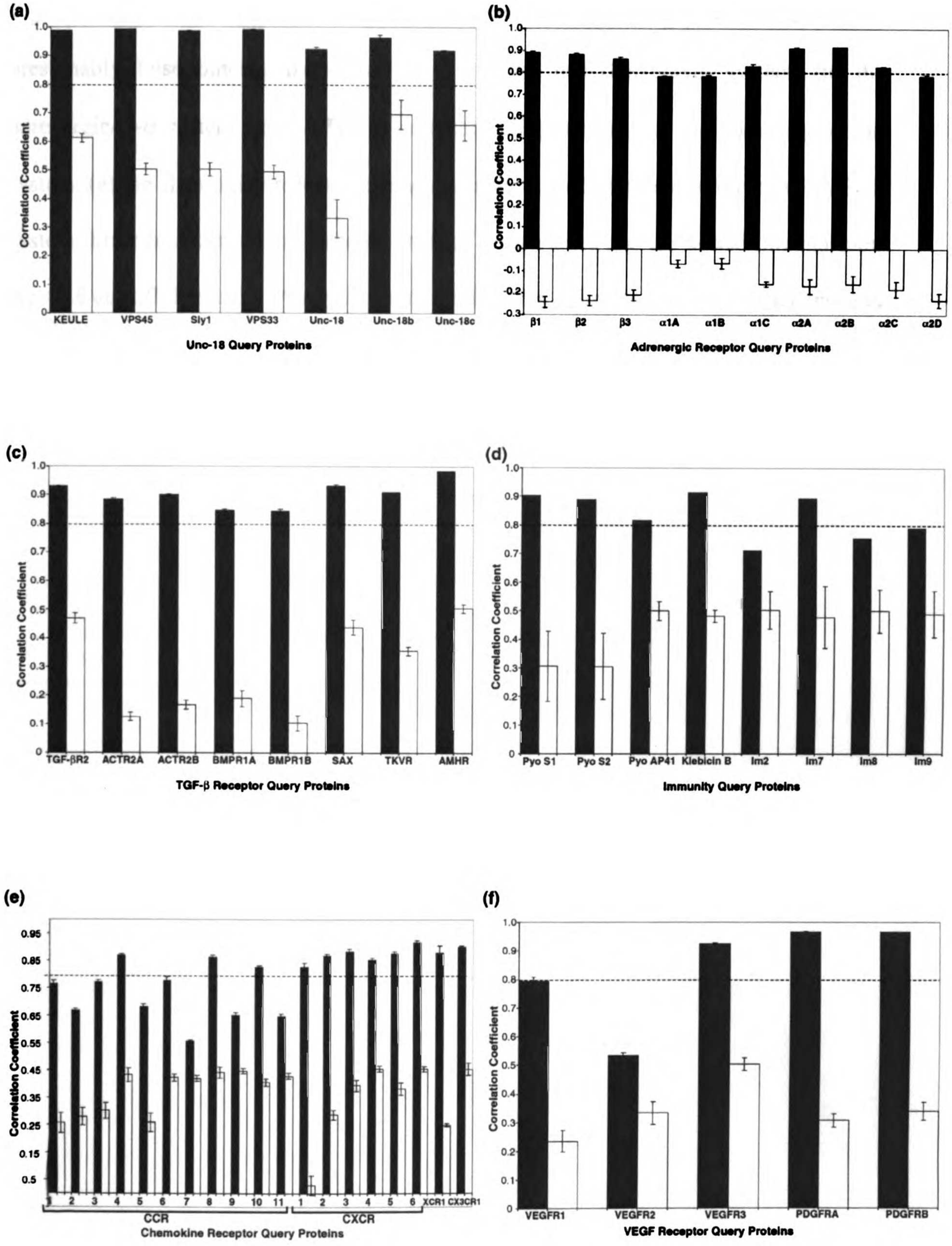
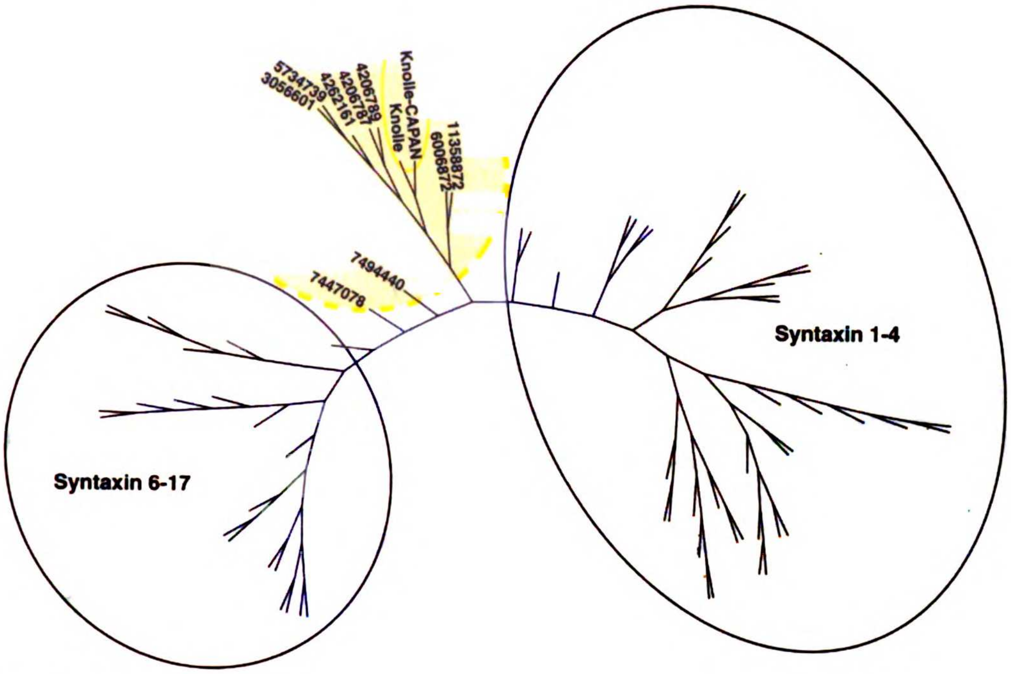

(a)
Correlation Coefficient
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
KEULE
VPS45
Sly1
VPS33
Unc-18
Unc-18b
Unc-18c
Unc-18 Query Proteins
(b)
Correlation Coefficient
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
β1
β2
β3
α1A
α1B
α1C
α2A
α2B
α2C
α2D
Adrenergic Receptor Query Proteins
(c)
Correlation Coefficient
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
TGF-βR2
ACTR2A
ACTR2B
BMPR1A
BMPR1B
SAX
TKVR
AMHR
TGF-β Receptor Query Proteins
(d)
Correlation Coefficient
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Pyo S1
Pyo S2
Pyo AP41
Klebicin B
Im2
Im7
Im8
Im9
Immunity Query Proteins
(e)
Correlation Coefficient
0.95
0.85
0.75
0.65
0.55
0.45
0.35
0.25
0.15
0.5
1
2
3
4
5
6
7
8
9
10
11
1
2
3
4
5
6
XCR1
CX3CR1
CCR
CXCR
Chemokine Receptor Query Proteins
(f)
Correlation Coefficient
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
VEGFR1
VEGFR2
VEGFR3
PDGFRA
PDGFRB
VEGF Receptor Query Proteins

Figure 4.1. The averaged correlation values of the known binding pairs (black) and the presumably false binding pairs (white) of (a) the unc-18/syntaxin system, (b) the adrenergic/G-α system, (c) the TGF-β receptor/TGF-β system, (d) the immunity/colicin system, (e) the chemokine receptor/chemokine system, and (f) the VEGF receptor/VEGF system. Error bars denote the standard error of these values and the dotted line indicates the >0.8 cut-off threshold. Analysis for all the families indicates the notable difference in correlation values between the true binding pairs and the potentially false binding pairs.

Syntaxin Family



Experimentally Determined Binding Specificity

Inferred Binding Specificity

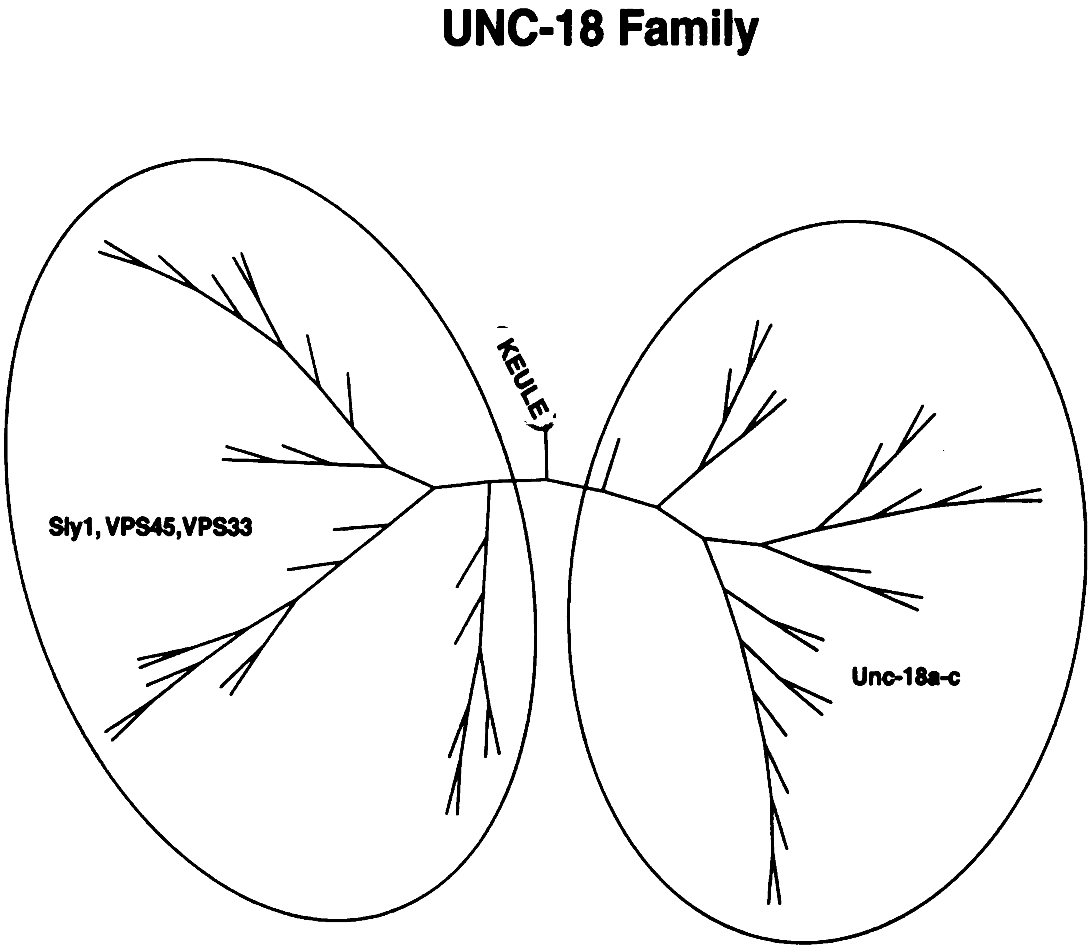
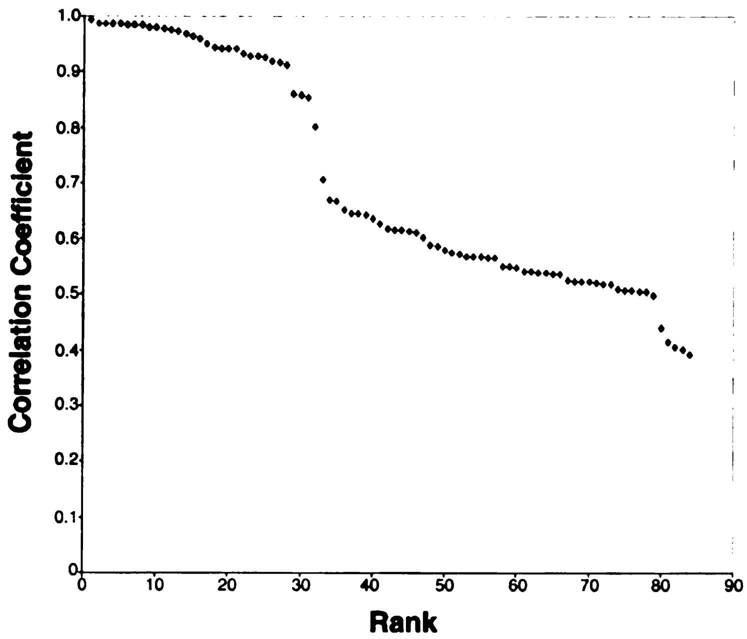


Figure 4.2. Phylogenetic trees of the (a) syntaxin family and the (b) unc-18 (Sec1) family. The area encompassed by the dotted circle indicates the search space for a potential KEULE binding partner through visual inspection. The solid circle outlines the known binding partner, Knolle, and its ortholog.

Identification of Potential KEULE Binding Partners



Rank	Partner	Correlation Value
1	GI #4262161	0.994
2	Knolle	0.987
3	Knolle-CAPAN	0.987
4	GI #3056601	0.986
5	GI #5734739	0.985
6	GI #7494440	0.984
7	GI #4206787	0.983
8	GI #7447078	0.980
9	GI #11358872	0.977
10	GI #4206789	0.976

Figure 4.3. Co-evolutionary analysis results for identification of potential KEULE binding partners. The second and third hit, indicated in red, are Knolle, the known binding partner, and its orthologue in *Capsicum annuum*. Some of the inferred binding partners with uncharacterized binding specificity, indicated in green, could also be potential KEULE interacting partners. The potentially false binding pairs are denoted in blue. These results suggest that a quantitative analysis of possible KEULE partners can focus the experimentalist on the most likely candidate, the second hit, instead of the set of 11 proteins found by visual inspection of the phylogenetic tree.

We believe that this quantitative approach will improve the chance of finding true binding pairs and reduce the chance of finding false pairs. Empirically, we have found two variations of this approach that are helpful in determining the number of pairs in the true positive group, the presumably false positive (unknown) group, and the orphan group. The first approach is to use a standard threshold of >0.8 to find binding pairs. The work in Goh *et al.*[11] and Pazos & Valencia (2001)[12] demonstrated that more than 66% of interacting domains could be detected in this way. Applying this algorithm, we were able to detect all 17 true known binding pairs from the syntaxin and unc-18 families out of a total of 296 possible binding pairs above the standard threshold value of >0.8 (Table 4.6). In addition, using this threshold criteria, we found a total of 18 binding pairs between previously characterized proteins that are currently not known to interact and 51 binding pairs that included orphans. At the most simple statistical level, the pre-test probability of finding a known binding pair would be 5.7% (there are 17 known binding pairs among a total of 296 protein pairs), a possibly false binding pair would be 35% (103/296), and an orphan binding pair would be 59% (176/296). Following the co-evolutionary analysis, the post-test probability of finding a known binding pair is 20% (17/86), a potentially false binding pair is 21% (18/86), and an orphan binding pair is 59% (51/86). Clearly, the set has become enriched in true known binding pairs and depleted of presumably false binding pairs. From the perspective of reducing the search space, the method refined the sample space so it still contained all (17/17) of the known binding pairs from the original sample, but only 17% (18/103) of the total probable false binding pairs, and 29% (51/176) of the total orphan binding pairs.

Frequently, experimental efforts are willing to characterize a small subset of alternative ligands in their search for a physiologically relevant binding partner. Consistent with this, our second approach was to choose the top 5 potential protein pairs. For these two families, most of the true binding pairs were found above the 0.9 correlation value (Figure 4.1), so the 0.8 cutoff for this family was generous and not as effective in reducing the number of potential false positives. By choosing the top 5 proteins for each query, the search space was further refined. The probability for finding a known binding pair was 37.5% (15/40), a possible false pair was 20% (8/40), and an orphan binding pair was 42.5% (17/40). Therefore compared to the original sample, the post-test probability of finding a known binding pair was increased by a factor of 6.6, and the probability of finding a possible false binding pair was decreased by 1.75 fold. Although the reduced search space did not contain all the known binding pairs (15/17) found in the original search space, it also did not retain as many presumably false binding pairs – 8% (8/103) of the unknown binding pairs from the original sample. 10% (17/176) of the orphan binding pairs from the original search space remained. Overall, all known binding pairs were found and 24% of the allegedly false binding pairs were retained at a correlation value of 0.71.

Adrenergic Receptors and G-proteinα-Subunits

The adrenergic receptors belong to the large superfamily of G-protein-coupled receptors (GPCRs). The binding of extracellular ligands to the GPCRs is thought to promote the receptor's association with distinct classes of G-proteins.[22-26] These G-proteins consist of α-subunits bound to βγ complexes attached to the plasma membrane.

Upon ligand activation, the receptor interacts with the heterotrimeric G-protein complex, which results in the dissociation of the G-α subunit from the βγ complex, triggering intracellular signaling cascades and a physiological response. In order to understand the physiological actions of a given GPCR, it is important to identify the specific G-proteins with which it interacts.

The adrenergic receptors belong to the biogenic amines receptor subfamily of GPCRs. This includes the seritonergic, dopaminergic, adrenergic, and muscarinic acetylcholine receptors. Studying this subfamily of receptors with their corresponding G-α subunits, we found that there was no correlation between the evolutionary histories of these two families. However, if the receptors were separated by their ligand specificity, then the correlations of the evolutionary histories of the receptors and their G-α subunits is statistically significant. This suggests that the GPCRs have a higher co-evolutionary signal for their ligands than they do for their corresponding G-α subunits. Nevertheless, because the GPCRs are known to bind to receptor-specific G-α subunits, the co-evolutionary signal can still be measured if the GPCRs are separated by their ligand specificity.

The adrenergic receptor family is illustrative of this observation. In our analysis, there were 142 sequences from the adrenergic receptor family, which included 10 characterized proteins with known G-α partners. We included 275 G-α subunit sequences from a variety of organisms. There was a total of 18 known G-α protein subunits found. Of these, 16 were known to bind to members of the adrenergic receptor family. The 10 adrenergic receptor proteins with known G-α partners were used as the query proteins to identify the interacting subunits (Table 4.2). The actual binding

partners within the two families had an overall correlation value of 0.76 while the averaged unknown binding pair correlation values ranged from –0.24 to 0, illustrating the large distinction between the true known binding pair correlations and the potentially false binding pair correlations (Figure 4.1b).

The original sample was divided as follows: 17% (38/220) known binding pairs, 42% (92/220) presumably false binding pairs, and 41% (90/220) orphan binding pairs. By using the >0.8 cutoff threshold, we identify 32 known binding pairs and no potentially false or orphan binding pairs (Table 4.6). By contrast, the top 5 strategy was less specific for these two families as most of the query proteins were known to bind to less than 5 proteins and the separation between the correlations of the known binding pairs and the presumably false binding pairs was so large. The probability of finding a known binding pair was 76% (38/50) and no false binding pairs were found. However, the probability of finding an orphan binding pair was 24% (12/50). Generally, this approach found all of the known binding pairs and only 13% (12/90) of the orphan binding pairs in the original sample. Both cases provided reasonable predictions and showed a tradeoff between sensitivity and specificity. Generally, for all queries at or above a correlation value of 0.7, all known binding partners were found, and less than 1% of the presumably false binding partners were found.

Table 4.2. Adrenergic Receptors and G-α Subunit Binding Pairs

G-α Subunits	Adrenergic Receptors
G_s	Adrenergic receptor β1, β2, β3
G_{olf}	Adrenergic receptor β1, β2, β3
G_{i1}, G_{i2}, G_{i3}	Adrenergic receptor α2A, α2B, α2C, α2D
G_{t1}, G_{t2}	Adrenergic receptor α2A, α2B, α2C, α2D
G_{gust}	Adrenergic receptor α2A, α2B, α2C, α2D
G_z	Adrenergic receptor α2A, α2B, α2C, α2D
G_{o1}, G_{o2}	Adrenergic receptor α2A, α2B, α2C, α2D
G_q	Adrenergic receptor α1A, α1B, α1C
G_{11}	Adrenergic receptor α1A, α1B, α1C
G_{14}	Adrenergic receptor α1A, α1B, α1C
G_{15}	Adrenergic receptor α1A, α1B, α1C
G_{16}	Adrenergic receptor α1A, α1B, α1C

These experimentally determined binding partners[26] were used to calculate the correlation coefficient and to infer additional binding pairs (see Methods).

TGF-β and TGF-β Receptors

Transforming Growth Factor β (TGF-β) superfamily members play a role in many cell processes, including early embryonic development, cell growth, differentiation, cell motility, and apoptosis.[27-29] TGF-β can dimerize to bind and activate a family of serine/threonine kinases known as the TGF-β receptor family. The TGF-β receptor family is divided into two subfamilies: type I receptors and type II receptors. TGF-β and its related factors activate signaling by binding and bringing together pairs of type I and type II receptors. Two modes of binding have been observed. Certain receptor type II homodimers must first bind the ligand before they can recruit the type I receptor into a complex.[30] This binding mode is characteristic of TGF-β and activin receptors.[31,32] In contrast, the BMP ligands show a higher binding affinity for the type I receptors than the type II receptors, although co-expression of both receptors has been shown to enhance the binding of the ligand.[33-36] The TGF-βs and their receptors were chosen to evaluate the co-evolution of proteins that utilized complex modes of binding.

In our co-evolutionary analysis, the ligands were assigned to bind to the receptor type for which they showed the highest affinity (Table 4.3). For the case of the TGF-β and activin receptors, only the type II receptors were designated to bind the ligand, with the implication that these ligands may co-evolve more strongly to their type II receptors than to their type I receptors. By contrast, for the BMP receptors, only the type I receptors were assigned to bind the ligand.

In this analysis, 348 sequences from the TGF-β family were used. Clustering species variants, these sequences identified a total of 55 proteins, 11 of which had known binding partners. Correspondingly, 203 receptor sequences drawn from 32 proteins were

included in the analysis. Of the 32 receptor proteins, 8 had known binding partners and were used as query proteins. The overall correlation for the experimentally determined binding partners within the two families was 0.7. The disparity between the values of the known binding pairs and the unknown are shown in Figure 1c, where all the queries for the known binding pairs had average correlation values >0.8 while the average correlation values of the presumably false binding pairs were <0.5.

By applying the >0.8 correlation threshold criteria, we were able to find 83% (15/18) of all known binding pair, while retaining 18% (13/70) of the binding pairs between characterized proteins that are not known to interact and 13% (45/352) of the orphan binding pairs. Statistically, the full search space of possible binding partners originally contained 4% (18/440) known binding pairs, 16% (70/440) unknown binding pairs, and 80% (352/440) orphan binding pairs. After the co-evolutionary analysis, the final sample consisted of 21% (15/73) known binding pairs, 18% (13/73) false binding pairs, and 62% (45/73) orphan binding pairs. Thus the post-test probability of finding a known binding pair was increased by a factor of 5.

Alternatively, the top 5 approach found 72% (13/18) of all known binding pairs, 4% (3/70) of all presumably false binding pairs, and 7% (24/352) of all orphan binding pairs. This resulted in a sample composed of 33% (13/40) known binding pairs, 7% (3/40) presumably false binding partners, and 60% (24/40) orphan binding pairs. For these two families, the top 5 strategy was able to better refine the search space, raising the probability of finding a known binding pair by a factor of 7.9 while lowering the probability of obtaining a false binding pair by 2.4 fold. Overall, at a correlation of 0.79

for all the queries, all of the known binding pairs were found while 7% of the presumably false binding pairs were found.

Table 4.3. TGF-β and TGF-β Receptor Binding Pairs

TGF-β Proteins	**TGF-β Receptors**
TGF-ßRII	**TGFß-1, TGFß-2, TGF-ß3**
ACTRIIA	**Activin βA, Activin βB**
ACTRIIB	**Activin βA, Activin βB**
AMHR	**MIS**
BMPRIA	**BMP-2, BMP-4, BMP-7, GDF-5**
BMPRIB	**BMP-2, BMP-4, BMP-7, GDF-5**
SAX	**Dpp**
TKVR	**Dpp**

These experimentally determined binding partners[28,75] were used to calculate the correlation coefficient and to infer additional binding pairs (see Methods).

DNase Colicins and their Immunity Proteins

Bacteria produce various antimicrobial molecules such as antibiotics, lytic enzymes, and bacteriocins that can kill other microbial competitors.[37] Bacteriocins are protein antibiotics that are generally effective against closely related species.[38] The colicins, an extensively studied group of bacteriocins produced by the *Escherichia coli* strains, have structural domains that perform different functions.[39] The N-terminal domain is implicated in translocation across the membrane of the target cell, the central domain is responsible for specific recognition of the target cell's extracellular surface receptor, and the C-terminal domain contains the toxic activity of the protein.[40]

The colicins are classified into three major groups according to their mode of action. The largest class is the pore-forming proteins, and the other two classes of colicins, the enzymatic E-colicins, have been identified as RNases or DNases. Each DNase colicin has its own specific immunity protein, which binds to the toxic domain of its cognate colicin and inhibits its cytotoxic activity while it is inside the producing cell.[41-43] The colicin-immunity protein recognition is highly specific,[44] since any given immunity protein will not, in general, provide protection for a non-cognate toxin. Therefore, we present the colicin DNase domains and their immunity proteins as a means to illustrate the correlated evolution of protein domains that interact with specific proteins.

The C-terminal domain, housing the cytotoxic activity of the E-colicins, has been shown to bind to its immunity protein.[45-47] A PSI-BLAST of the C-terminal domain of the DNases retrieved nine sequences consisting of eight colicins and one orphan protein, a Usp protein found in *E. coli*. A similar search of the DNase immunity proteins found eight sequences of the eight corresponding immunity proteins. The immunity proteins were used as the query proteins to find their specific binding partner. The overall correlation for the previously determined binding parters within the two families was 0.67. All the queries had average correlation values for known binding pairs of >0.7 and average correlation values for presumably false binding pairs of <0.5 (Figure 4.1d). Although the top hit was the correct partner for each of the Im2, Im8, and Im9 queries, the averaged correlation values for the known binding partner for Im2, Im8, and Im9 were lower than the rest of the queries in that family, suggesting a weaker co-evolutionary signal towards its cognate protein compared to the rest of the queries.

However, experimental studies have confirmed that although each immunity protein has a higher affinity for its cognate colicin, there is detectable cross-reactivity between ColE9 and the Im8 and Im2 proteins.[48]

Originally, the complete sequence set included 12% (8/64) true binding pairs and 88% (56/64) potentially false binding pairs. After applying the co-evolutionary analysis using the >0.8 correlation threshold, the search space contained 71% (5/7) true binding pairs and 29% (2/7) false binding pairs, illustrating the 18-fold enrichment of true binding pairs in the search space. The method was able to detect 63% (5/8) of all known binding pairs and 4% (2/56) of all possible false binding pairs. By picking the top 5 pairs, the final search space was made up of 20% (8/40) true positives and 80% (32/40) false positives. This sample contained all of the known binding pairs but also 57% (32/56) of the false binding pairs. Because each immunity protein only has one cognate binding partner, picking the top five pairs necessarily leads to more false positive identifications. Since neither of these approaches found a pair containing the orphan protein, Usp, it could be inferred that Usp is not a good candidate binding partner for any of the known immunity proteins. Generally, for all queries at a correlation threshold of 0.71, this analysis was able to find all of the known binding pairs while retaining 16% of the presumably false positives.

Table 4.4. Chemokine and Chemokine Receptor Binding Pairs

Chemokines	Chemokine Receptors
CCR1	MIP-1α, RANTES, MCP-3, HCC-1, HCC-4, MPIF-1
CCR2	MCP-1, MCP-2, MCP-3, MCP-4, MCP-5, HCC-4
CCR3	MCP-2, MCP-4, RANTES, EOTAXIN, EOTAXIN-2, EOTAXIN-3
CCR4	TARC, MDC
CCR5	MIP-1α, MIP-1β, RANTES
CCR6	MIP-3α
CCR7	MIP-3β, SLC
CCR8	I-309
CCR9	TECK
CCR10	CTACK, MEC
CCR11	TECK, SLC, MIP-3β
CXCR1	IL-8, GCP-2, GROα
CXCR2	IL-8, GCP-2, GROα, GROβ, GROγ, ENA-78, NAP-2
CXCR3	MIG, IP-10, I-TAC
CXCR4	SDF-1
CXCR5	BLC
CXCR6	CXCL16
XCR1	XCL1, XCL2
CX3CR1	CX3C

These experimentally determined binding partners[76,77] were used to calculate the correlation coefficient and to infer additional binding pairs (see Methods).

Chemokine and Chemokine Receptor Families

Chemokines are a family of chemotactic cytokines that activate transmembrane G-protein coupled receptors (GPCRs) on the cell surface to regulate diverse biological processes which include leukocyte trafficking, angiogenesis, hematopoiesis, and organogenesis.[49,50] A single chemokine can bind to more than one receptor and a given

receptor can bind to several chemokines (Table 4.4). This can pose a challenge for investigators who wish to elucidate the physiological activities of chemokines *in vivo*.[51]

Despite the fact that both receptors and chemokines can bind multiple partners with high affinity, the two families do co-evolve and have an overall correlation value of 0.66 for all known binding partners. This is slightly higher than the 0.57 correlation previously reported.[11] This increased value is probably due to the greater number of binding partners determined since the previous report. The chemokines and their receptors were chosen to represent the co-evolution of a standard ligand-receptor system.

In total, there are 147 sequences in the chemokine family comprising of 42 proteins with previously characterized binding specificities and seven orphan proteins. Additionally, 19 characterized proteins made up of 102 sequences from 28 different organisms in the chemokine receptor family were used. The 19 proteins in the chemokine receptor family were each utilized as query proteins in order to test the predictive results of this algorithm. Over half of the protein queries have known binding pair coefficient values >0.8 and all of the protein queries have false binding pair coefficient values <0.5 (Figure 4.1e). In the case of the chemokine receptors CCR7, CCR9, and CCR11, the average correlation coefficients for their true binding pairs were lower than the rest of the chemokine receptors because they all shared the same binding partners in a non-uniform manner. CCR7 was known to interact with SLC and MIP-3β, while CCR9 was known to bind to TECK. However, CCR11 was known to bind to SLC, MIP-3β, and TECK. The algorithm was meant to find close sequence neighbors that share the same binding partners. However, since these receptors share some but not all

binding partners, the co-evolutionary analysis was less able to accurately infer all the binding pairs.

Before the analysis, the distribution included 6% (52/931) known binding pairs, 80% (746/931) presumably false binding pairs, and 14% (133/931) orphan binding pairs (Table 4.6). Applying the >0.8 threshold value yielded a post-test distribution where 37% (30/81) of the results were known binding pairs, 31% (25/81) were supposedly false binding pairs, and 32% (26/81) involved binding pairs containing orphans. We found an increased probability of finding an orphan pair, suggesting additional binding information for previously uncharacterized proteins. The final sample was composed of 58% (30/52) of the total known binding pairs from the original search space, with only 3% (25/746) of the presumably false binding pairs, and 20% (26/133) of the orphan binding pairs from the starting set.

By picking the top 5 protein pairs, the probability of finding a known binding pair was reduced to 30% (27/91), while there remained a 35% (32/91) chance of finding a potentially false binding pair and 35% (32/91) chance of finding an orphan binding pair. The final search space was reduced to 52% (27/52) of the total known binding pairs, 4% (32/746) of the total possible false binding pairs, and 24% (32/133) of the orphan binding pairs. For this family, the 0.8 cutoff value approach yielded better results because of the lower overall correlation value for these two families and the extra noise gathered from queries with less than 5 known binding partners. An overall analysis of the results showed that at a correlation threshold of 0.54, the method found all of the known binding pairs while retaining 27% of all binding pairs with characterized proteins not known to interact.

VEGF and VEGF Receptors

Vascular endothelial growth factor (VEGF) is a member of the cystine-knot family[52] that regulates multiple biological functions such as endothelial cell differentiation (vasculogenesis) and formation of new capillaries from pre-existing vessels (angiogenesis) during development.[53] Angiogenesis is not only involved with physiological processes such as wound healing, but is also involved in pathological processes such as tumor growth, diabetic retinopathy, and rheumatoid arthritis.[54] Obtaining a better understanding of the molecular mechanisms underlying the biological function of VEGF and the angiogenic response may lead to new therapeutic approaches.

From PSI-BLAST, we obtained 94 sequences from the VEGF family, which encoded seven genes across approximately 15 species, and 5 genes from the VEGF receptor family with species variants totaling 80 sequences. The five receptors from the receptor family were used as query proteins to infer their binding partners (Table 4.5). The overall correlation for the known binding partners within the two families was 0.65. All but one of the queries had average correlation values with their known binding pairs of >0.75 and average correlation values with presumably false binding pairs of <0.5 (Figure 4.1f). The VEGFR2 (KDR) receptor was the only query protein with an average correlation value with known binding pairs below 0.75. VEGFR1 (Flt-1) receptor, the closest sequence neighbor to VEGFR2 (KDR) receptor, has been shown to bind to VEGFA, VEGFB, and Plgf-1 (placental growth factor-1). Although VEGFR2 (KDR) receptor also binds VEGFA, it has not been shown to bind either VEGFB or Plgf-1. This can lead to a lower averaged correlation value for its true binding partners.

At the outset, 23% (8/35) of the possible pairs were known to bind and 77% (27/35) of the possible pairs were presumed not to bind. After applying the co-evolutionary analysis with a >0.8 cut-off threshold, the resulting sample contained 70% (7/10) true binding pairs and 30% (3/10) presumably false binding pairs. The reduced sample contained 88% (7/8) of the known binding pairs and 11% (3/27) of the presumably false binding pairs. By picking the pairings with the top 5 scores, the resulting sample consisted of 32% (8/25) known binding partners and 68% (17/25) presumably false binding partners. This sample contained all (8/8) of the known binding pairs and 63% (17/27) of the presumably false binding pairs. This high potential false positive was due to the fact that most of the receptors had only one or two known binding partners.

Table 4.5. VEGF and VEGF Receptor Binding Pairs

VEGF Proteins	VEGF Receptors
VEGFR1	**VEGF-A, VEGF-B, PIGF**
VEGFR2	**VEGF-A**
VEGFR3	**VEGF-C, VEGF-D**
PDGFRA	**PDGF-A**
PDGFRB	**PDGF-B**

These experimentally determined binding partners[78,79] were used to calculate the correlation coefficient and to infer additional binding pairs (see Methods).

Identification of Potential Binding Pairs for Uncharacterized Proteins

A significant number of the surveyed families including the syntaxin/unc-18 and the TGF-β/TGF-β receptor systems included orphan ligands with significant correlated co-evolution with a known receptor. The results of the analysis suggested possible

binding partners for several orphans in these families. For example, the query for the KEULE protein resulted in a correlation value of 0.986 when paired with Syr1, far higher than when paired with other orphans, suggesting Syr1 as a possible interacting partner for KEULE. Recent studies indicate that there may be another protein, other than Knolle, that interacts with KEULE. One observation in plants with mutations in KEULE is that they lack root hairs. However, Knolle mutants do not show this same effect. One explanation for this observation is that KEULE can interact with another protein as well as Knolle.[55] Additionally, the Syr1 protein has been found in roots.[56] This would be appropriate for a KEULE-interacting protein that could generate specialized tissue in plants.

Another orphan, syntaxin 16, had correlation values above 0.9 for both the Vps33 and the Vps45 proteins; and the syntaxin 12 protein had correlation values above 0.95 for the Vps33, Vps45, and Sly1 proteins (Figure 4.4a). Incorporating known tissue localization information, we can infer that syntaxin 16 may be a more likely binding partner of Vps45 than Vps33. Syntaxin 16 is located in the *trans*-Golgi network (TGN), similar to syntaxin 6, a binding partner for the Vps45 protein.[17] Similarly, syntaxin 12, a protein found in the endosome, may be a more likely binding partner for Vps33, since syntaxin 7, a binding protein to Vps33 is also found in the endosome.

The BMP type 1 receptor protein queries found four proteins that had average correlation scores >0.75. The co-evolutionary analysis revealed that GDF7 had a 0.85 and 0.83 correlation value for BMP receptor 1A and BMP receptor 1B, respectively, indicating a possible interaction between GDF7 and the BMP type 1 receptors.

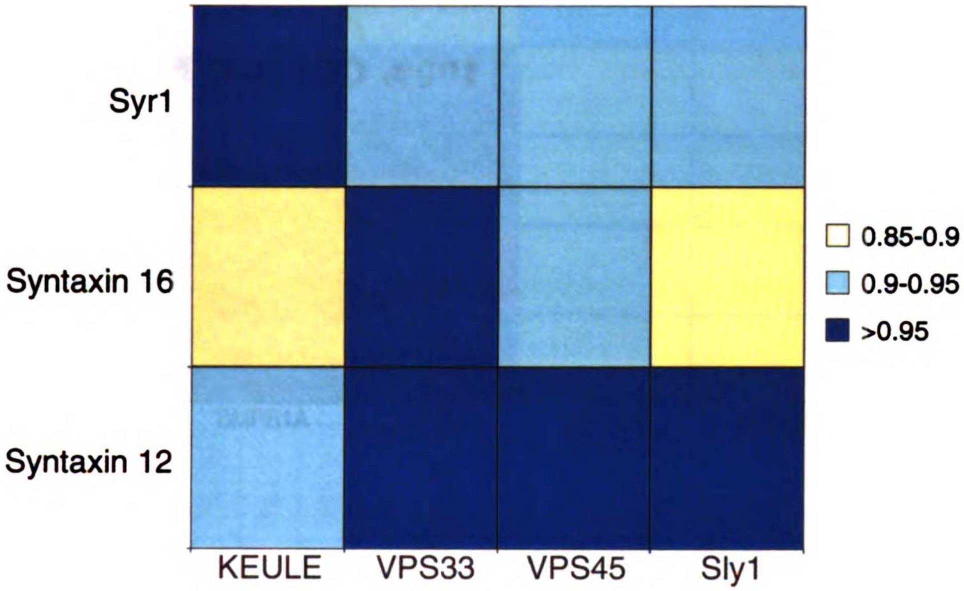
Observations during seminal vesicle development indicate that GDF7 was shown to be expressed in the mesenchyme at the same time as the two BMP type 1 receptors were in the epithelium. This is consistent with a role for GDF7 in mesenchymal-epithelial interactions.[57] Co-evolutionary analysis revealed that GDF6 had a correlation value of 0.77 and 0.74 for the two BMP type 1 receptors. Recent data showed that GDF6 could form heterodimers and be co-expressed with BMP2, a ligand for the BMP type 1 receptors.[58] This suggests that GDF6 could be a ligand for the BMP type 1 receptors. Other potential ligands for the BMP type 1 receptors include BMP5 and BMP6 which both have correlation values above 0.75.

Co-evolutionary analysis of the activin type 2 receptors suggested two additional binding partners, GDF8 and GDF11, with correlation values above 0.65 (Figure 4.4b). A recent experimental study supports this computational result by demonstrating the *in vitro* binding of GDF8 to both activin type 2 receptors.[59] These studies also suggest possible *in vivo* interactions between GDF8 and the activin type 2 receptors. Lee & McPherron[59] observed that the expression of the dominant-negative form of the activin type 2B receptor caused an increase in muscle mass similar to what was found in GDF8 knockout mice. Additionally, the mutant phenotype of GDF11, including additional thoracic vertebrae and kidney defects, has similarly been observed in mice lacking the activin type 2B receptor.[60] Together, these data provide corroborative evidence that GDF8 and GDF11 could be ligands for the activin type 2 receptors.

For known ligand-receptor systems, the co-evolutionary analysis suggests biologically plausible binding receptors for certain orphan ligands. Notably, the pair of binding receptors with the highest affinity is inferred to bind the ligand, offering

supportive evidence for mechanisms of protein interaction. However, while members of the ligand family may bind as hetero- or homodimers, different combinations of receptor-receptor complexes can also lead to differences in biological phenotype from previously characterized complexes, adding additional complexity to the study of this system.

Averaged Correlations of Syntaxin/Unc-18 Inferred Binding Pairs



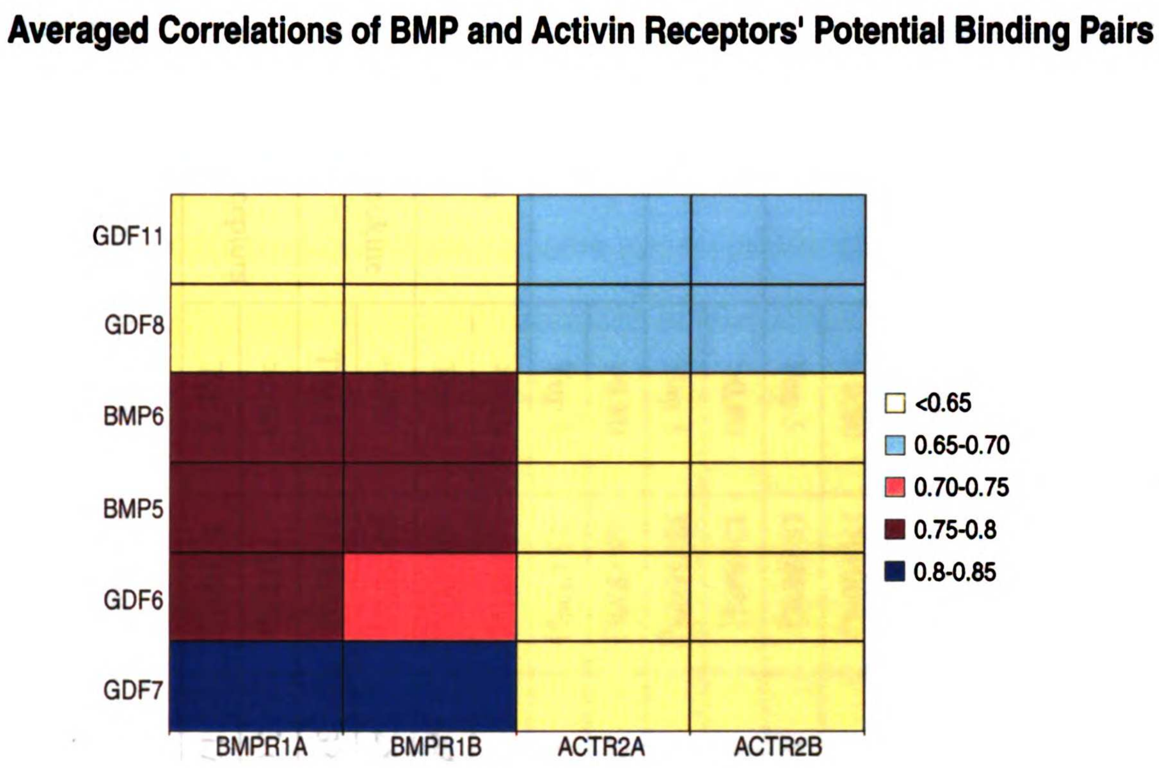


Figure 4.4. Averaged correlation values of inferred binding pairs comprised of orphan proteins and known (a) unc-18 proteins and (b) TGF-β receptors. The range of correlation scores are represented by colors to show the differentiation of suggested binding specificity. For example, Syr1 appears to be a relatively specific partner for KEULE, while syntaxin 12 could be a more promiscuous ligand and bind VPS33, VPS45, and Sly1.

Table 6: Number of True Binding Partners Found at Selected Correlation Value

Protein Families	Correlation Value Criteria	True Predicted Binding Pairs*	Predicted Binding Pairs of Uncertain Significance+	Predicted Binding Pairs of Orphans#	Total Number of Known Binding Pairs	Total Number of Possible Pairs (without orphans)	Total Number of Possible Pairs (includes orphans)	Overall Correlation
Unc-18/Syntaxin	> 0.80	17(100%)	18 (17%)	51 (29%)	17	120 (8x15)	296 (8x37)	0.82
	Top 5	15 (88%)	8 (7.8%)	17 (9.7%)				
Adrenergic Receptors/Gα	>0.80	32 (84%)	0 (0%)	0 (0%)	38	130 (10x13)	220 (10x22)	0.76
	Top 5	38 (100%)	0 (0%)	12 (13%)				
TGF-ß/TGF-β Receptors	>0.80	15 (83%)	13 (18%)	45 (13%)	18	88 (8x11)	440 (8x55)	0.70
	Top 5	13 (72%)	3 (4.3%)	24 (6.8%)				
Immunity/Colicin DNAse Domains	>0.80	5 (63%)	2 (3.6%)	0 (0%)	8	64 (8x8)	64 (8x8)	0.67
	Top 5	8 (100%)	32 (57%)	0 (0%)				
Chemokine/Chemokine Receptors	>0.80	30 (58%)	25 (3.3%)	26 (20%)	52	798 (19x42)	931 (19x49)	0.66
	Top 5	27 (52%)	32 (4.3%)	32 (24%)				
VEGF/VEGF Receptors	>0.80	7 (88%)	3 (11.1%)	0 (0%)	8	35 (5x7)	35 (5x7)	0.65
	Top 5	8 (100%)	17 (63%)	0 (0%)				

* Cell Value/Total Number of Known Binding Pairs

+ Cell Value/(Total Number of Possible Pairs without orphans -Total Number of Known Binding Pairs)

Cell Value/(Total Number of Possible Pairs including orphans-Total Number of Possible Pairs without orphans)

DISCUSSION

Co-evolutionary analysis exploits the notion that correlated divergent evolution in families of structurally homologous proteins can be used to identify binding partners between the two families. Since proteins that interact should co-evolve in order to maintain the energetically and structurally relevant features of the binding interface, variations in the sequence could influence their binding specificity. By relating the sequence similarity of a set of proteins to their binding partner preferences, the binding specificity of an uncharacterized protein can be inferred by its sequence similarity to other characterized proteins within the same family. Results of the analysis show that binding partners can be quantitatively identified for proteins in diverse homologous protein families with approximately 58–100% sensitivity (predicted true binding pairs/all true binding pairs) and 82–100% specificity (1-(predicted false binding pairs/all false binding pairs)) for a threshold value >0.8. This greatly reduces the search space for finding potential binding partners in an objective manner and serves as a useful algorithm that can be readily applied to many protein systems.

Refining the co-evolutionary information can lead to a better quantitative measure of co-evolution between two protein families. In this study, we show how co-evolutionary signals of interacting multi-domain proteins can be isolated and, thus, contribute to a more accurate analysis and better overall correlation between two protein families. One example occurs in proteins containing modular domains that each have separate functions and may interact with different proteins. In the colicin/immunity protein system, the colicin protein is made up of three domains – an N-terminal

translocation domain, a central receptor-binding domain, and a C-terminal cytotoxic domain.[40] The overall correlation of the full colicin protein to its immunity protein was 0.51. However, the overall correlation of the C-terminal DNase domain of the colicin protein with its immunity protein was 0.67, suggesting a more recognizable co-evolution of the immunity protein to the colicin DNase domain than to the whole colicin protein. Since the other two domains of the colicin are known to bind to other proteins and not to the immunity proteins,[45,47,61] the inclusion of these domains in the analysis was potentially confounding and led to a lower overall correlation value between the interacting proteins. In the reverse fashion, for proteins with domains of uncharacterized function, this analysis can be used to determine which domains can interact with a known binding protein.

Another factor that influences the accuracy of the co-evolutionary analysis is based on the fact that most protein families interact and co-evolve with several different families. This can lead to ambiguity in the co-evolutionary analysis if a certain protein family (Family A) binds to a protein family (Family C) but co-evolves more strongly to another protein family (Family B). By dismissing the evolutionary information from Family B, the co-evolutionary analysis between Family A and Family C becomes more precise. This approach could be useful in systems containing non-modular multi-domain proteins. In the case of the GPCRs, the receptors show sequence similarities that correspond to the binding specificities of their ligands rather than to their corresponding G-α protein subunits. The ligand-specific information was removed by partitioning the GPCRs into ligand-specific subclasses. Subsequently, the co-evolution of the receptors for their G-α protein subunits could then be more easily measured, resulting in a highly

accurate prediction of 83% (sensitivity) of known binding pairs without any false positives (100% specificity) for pairs with a > 0.8 correlation value.

The general predictive power of this method is partially reflected in the overall correlation value but other factors, such as promiscuity and the type of interaction complex formed, contribute to the overall prediction results. In general, the averaged correlation values of true known binding pairs are clearly higher than those of the presumably false binding pairs. Certain queries do not show as strong a distinction as others. As mentioned in the results, the VEGFR2 receptor and some of the chemokine receptors had a lower averaged correlation value with their known binding pairs due, in part, to the non-uniformity of binding specificities, where close sequence neighbors shared some but not all experimentally known binding partners. However, some predicted unknown binding pairs could be instances of the fact that cross-reactivity can occur for non-cognate binding partners. For example, some of the immunity proteins, which only have one known cognate binding partner, had predicted binding pairs in the top 5 that were not cognate partners, but had been experimentally shown to bind with a lower binding affinity.[48]

The mode of binding for certain systems can also be a determinant that influences the predictive power of the co-evolutionary analysis. For most systems that form a 1:1 complex, this becomes less of a factor. In the circumstance where the complex is made up of more than two different proteins, if two proteins show a higher binding affinity for each other than the other proteins in the complex, we have shown that the co-evolutionary analysis can still provide rational and sensible inferences for these binding pairs. For TGF-β, which binds to two different receptors simultaneously, we found that

subsetting the ligands into type I or type II based on their receptor binding preferences led to better statistical separation of the known binding pairs from the total set.

In this paper, we have described two types of criteria for analyzing the results of the co-evolutionary analysis algorithm that illustrate the trade-off between sensitivity and specificity. In general, for protein families such as the immunity protein family or VEGF receptor family, which on average bind to one or two specific proteins, the 0.8 cut-off can provide a high likelihood ratio for finding the true binding pairs, because the true predictions will usually cluster at the top scores with a large differentiation between the correlations of the true binding partners and the false binding partners. By contrast, picking the top five pairs from the results may introduce a larger number of false positives because of the small number of actual binding partners for each protein. For families that contain proteins that are known to bind promiscuously to three or more proteins in the corresponding family, picking the top 5 will generally produce a better set of candidates for finding interacting partners. The results from this analysis illustrate the high likelihood ratios (7.75 and above) for obtaining the true binding pairs.

The co-evolutionary analysis is based on the hypothesis of divergent evolution of homologous proteins, and anchored by the availability of specific binding information to make accurate inferences about orphan protein binding pairs. In the cases of convergent evolution, where ligand A evolved to bind to a receptor whose closest sequence neighbor binds to a ligand that is structurally different from ligand A, the co-evolutionary analysis should not be able to provide conclusive results and is likely to be misleading. Also, certain predictions may actually have reasonable binding affinity in vitro but may never occur in vivo due to tissue localization and gene expression patterns. This could explain

why certain proteins have the same binding interface, but can bind different partners. Incorporating spatial and temporal localization information with co-evolutionary analysis could result in a more precise inference of biological function. Other causes of false positives or negatives, such as sequence misalignment events that lead to an inaccurate distance matrix, would further skew prediction results. If the sequence similarity is extremely low, then the small probability of a useful alignment can undermine the accuracy of our method. In addition, the utility of the derived sequence homology score as a surrogate for the evolutionary distance between the sequences can greatly influence the predictive power of the co-evolutionary analysis.

CONCLUSIONS

The co-evolutionary analysis of various protein systems supports the hypothesis that proteins that interact also must co-evolve. Through the use of sequence information, this method can quantitatively measure the co-evolution between two proteins and, therefore, make specific and impartial inferences about plausible protein-protein interactions. We applied this analysis to identify candidate binding partners for orphan ligands by reducing the search space to a small subset of proteins enriched with likely binding partners. By providing a quantitative measure for inferring binding partners, potential interacting proteins in various protein-protein interaction systems can now be identified in a fast and objective manner.

This approach will become even more powerful as more genes are cloned, thereby filling in the gaps of missing sequence information. While maximizing the correlation values of proteins between two families could provide a pure computational "two-hybrid", we believe that increasing amounts of experimental information can only improve the utility of the co-evolutionary strategy. Identifying and analyzing only the specific residues important in maintaining the interactions between two proteins could further improve and refine the analysis. Operationally, we have found it difficult to extract a reliable evolutionary signal from small subsets of the sequence such as the residues in the binding interface. In its current form, the co-evolutionary analysis is a robust and versatile approach to infer which proteins and/or domains are likely to interact when they are parts of larger families that are known to interact.

METHODS

Sequence Analysis

Sequences in each protein family were retrieved using PSI-BLAST[62] and the complete non-redundant database updated as of January 2002. Sequence distance information from the orthologs of proteins with known binding specificities was also used to build the distance matrices. PSI-BLAST was run with the default parameters (0.001 e-value cutoff for inclusion of a sequence in the matrix calculations, filtering turned on, and a maximum of 3 rounds). Multiple sequence alignments and pairwise distances generated from the multiple sequence alignments were constructed from ClustalW.[13]

Correlation Analysis

Distance matrices were constructed in the order of known experimentally determined binding pairs as described previously.[11] Therefore, for a given receptor family, we constructed a receptor distance matrix X, where X is defined as an N x N matrix and N is equal to the number of known binding receptor-ligand interactions. For the corresponding ligand family, we constructed a similar distance matrix Y. Therefore, an entry, X_{ij}, is the pairwise distance between sequence m_i and sequence $m_{j,}$ in the receptor family. Y_{ij}, in the matrix Y, signifies the pairwise distance between sequence n_i and sequence n_j in the ligand family. Sequence n_i is experimentally known to bind to sequence m_i and sequence n_j is known to bind to sequence m_j.

In order to test the robustness and accuracy of this method, we removed all known binding data for each receptor in the receptor families and then used that receptor as a query protein. Receptor families were chosen as query families because they usually contain fewer proteins than their corresponding ligands. For example, for the syntaxin system, the query family was the unc-18 family (also known as Sec1); and for the colicin system, the immunity protein family was the query family.

For each query/uncharacterized protein in the receptor family found in the multiple sequence alignment, we extended the receptor distance matrix X by adding an additional row and column of its pairwise distances to every other receptor in the existing receptor matrix. Similarly, we extended the ligand distance matrix Y by adding a new column and row vector of pairwise distances between a protein in the ligand multiple sequence alignment to every other ligand in the existing ligand matrix. This results in two $(N+1)$ *x* $(N+1)$ matrices. The linear correlation coefficient (r) of these two row vectors is calculated according to the standard equation:[63]

$$r = \frac{\sum_{i=1}^{N+1}(X_{(N+1)i} - \overline{X}_{(N+1)})(Y_{(N+1)i} - \overline{Y}_{(N+1)})}{\sqrt{\sum_{i=1}^{N+1}(X_{(N+1)i} - \overline{X}_{(N+1)})^2}\sqrt{\sum_{i=1}^{N+1}(Y_{(N+1)i} - \overline{Y}_{(N+1)})^2}}$$

with $-1 \leq r \leq 1$ where $\overline{X}_{(N+1)}$ is the mean of all distances in the row vector $X_{(N+1)}$ and $\overline{Y}_{(N+1)}$ is the mean of all distances in the row vector $Y_{(N+1)}$.

The new row vector of pairwise distances for the query protein in the receptor family is kept fixed while row vectors for each protein in the ligand multiple sequence

alignment are generated. A correlation coefficient is generated for the query protein row vector and every protein row vector in the ligand multiple alignment. A total of K number of correlations result for each query protein, where K is the number of total proteins in the ligand multiple sequence alignment.

ACKNOWLEDGEMENTS

C-S.G. was supported by the UCSF Interface of Sciences Fellowship awarded by the Burroughs Wellcome Fund. This work was supported by grants from the NIH (to F.E.C.). We thank Anthony Lau and David M. Kingsley for helpful discussions.

REFERENCES

1. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-6 (1989).

2. Neubauer, G., Gottschalk, A., Fabrizio, P., Seraphin, B., Luhrmann, R. & Mann, M. (1997). Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc Natl Acad Sci U S A*, **94**, 385-90.

3. Mendelsohn, A.R. & Brent, R. Protein interaction methods--toward an endgame. *Science* **284**, 1948-50. (1999).

4. Mann, M., Hendrickson, R.C. & Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem* **70**, 437-73 (2001).

5. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., *et al.* (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180-3.

6. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23**, 324-8. (1998).

7. Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90 (1999).

8. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751-3.

9. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-86 (1999).

10. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **96**, 4285-8 (1999).

11. Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D. & Cohen, F.E. Co-evolution of proteins with their interaction partners. *J Mol Biol* **299**, 283-93. (2000).

12. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng* **14**, 609-14. (2001).

13. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673-80 (1994).

14. Sollner, T., Bennett, M.K., Whiteheart, S.W., Scheller, R.H. & Rothman, J.E. A protein assembly-disassembly pathway in vitro that may correspond to sequential steps of synaptic vesicle docking, activation, and fusion. *Cell* **75**, 409-18. (1993).

15. Bock, J.B., Matern, H.T., Peden, A.A. & Scheller, R.H. A genomic perspective on membrane compartment organization. *Nature* **409**, 839-41. (2001).

16. Chen, Y.A. & Scheller, R.H. SNARE-mediated membrane fusion. *Nat Rev Mol Cell Biol* **2**, 98-106. (2001).

17. Teng, F.Y., Wang, Y. & Tang, B.L. The syntaxins. *Genome Biol* **2**(2001).

18. Hata, Y., Slaughter, C.A. & Sudhof, T.C. Synaptic vesicle fusion complex contains unc-18 homologue bound to syntaxin. *Nature* **366**, 347-51. (1993).

19. Misura, K.M., Scheller, R.H. & Weis, W.I. Three-dimensional structure of the neuronal-Sec1-syntaxin 1a complex. *Nature* **404**, 355-62. (2000).

20. Carr, C.M., Grote, E., Munson, M., Hughson, F.M. & Novick, P.J. Sec1p binds to SNARE complexes and concentrates at sites of secretion. *J Cell Biol* **146**, 333-44. (1999).

21. Hanson, P.I. Sec1 gets a grip on syntaxin. *Nat Struct Biol* **7**, 347-9. (2000).

22. Dohlman, H.G., Thorner, J., Caron, M.G. & Lefkowitz, R.J. Model systems for the study of seven-transmembrane-segment receptors. *Annu Rev Biochem* **60**, 653-88 (1991).

23. Conklin, B.R. & Bourne, H.R. Structural elements of G alpha subunits that interact with G beta gamma, receptors, and effectors. *Cell* **73**, 631-41. (1993).

24. Neer, E.J. Heterotrimeric G proteins: organizers of transmembrane signals. *Cell* **80**, 249-57. (1995).

25. Bourne, H.R. How receptors talk to trimeric G proteins. *Curr Opin Cell Biol* **9**, 134-42. (1997).

26. Wess, J. Molecular basis of receptor/G-protein-coupling selectivity. *Pharmacol Ther* **80**, 231-64. (1998).

27. Derynck, R. & Feng, X.H. TGF-beta receptor signaling. *Biochim Biophys Acta* **1333**, F105-50. (1997).

28. Massague, J. TGF-beta signal transduction. *Annu Rev Biochem* **67**, 753-91 (1998).

29. Piek, E., Heldin, C.H. & Ten Dijke, P. Specificity, diversity, and regulation in TGF-beta superfamily signaling. *Faseb J* **13**, 2105-24. (1999).

30. Wrana, J.L., Attisano, L., Wieser, R., Ventura, F. & Massague, J. Mechanism of activation of the TGF-beta receptor. *Nature* **370**, 341-7. (1994).

31. Boyd, F.T. & Massague, J. Transforming growth factor-beta inhibition of epithelial cell proliferation linked to the expression of a 53-kDa membrane receptor. *J Biol Chem* **264**, 2272-8. (1989).

32. Laiho, M., Weis, M.B. & Massague, J. Concomitant loss of transforming growth factor (TGF)-beta receptor types I and II in TGF-beta-resistant cell mutants implicates both receptor types in signal transduction. *J Biol Chem* **265**, 18518-24. (1990).

33. Gilboa, L., Nohe, A., Geissendorfer, T., Sebald, W., Henis, Y.I. & Knaus, P. (2000). Bone morphogenetic protein receptor complexes on the surface of live cells: a new oligomerization mode for serine/threonine kinase receptors. *Mol Biol Cell*, **11**, 1023-35.

34. Kirsch, T., Nickel, J. & Sebald, W. BMP-2 antagonists emerge from alterations in the low-affinity binding epitope for receptor BMPR-II. *Embo J* **19**, 3314-24. (2000).

35. Liu, F., Ventura, F., Doody, J. & Massague, J. Human type II receptor for bone morphogenic proteins (BMPs): extension of the two-kinase receptor model to the BMPs. *Mol Cell Biol* **15**, 3479-86. (1995).

36. Rosenzweig, B.L., Imamura, T., Okadome, T., Cox, G.N., Yamashita, H., ten Dijke, P., *et al.* (1995). Cloning and characterization of a human type II receptor for bone morphogenetic proteins. *Proc Natl Acad Sci U S A*, **92**, 7632-6.

37. James, R. (ed.) *Bacteriocins, microcins and lantibiotics*, (Springer-Verlag, 1992).

38. Baba, T. & Schneewind, O. Instruments of microbial warfare: bacteriocin synthesis, toxicity and immunity. *Trends Microbiol* **6**, 66-71. (1998).

39. Ohno-Iwashita, Y. & Imahori, K. Assignment of the functional loci in colicin E2 and E3 molecules by the characterization of their proteolytic fragments. *Biochemistry* **19**, 652-9. (1980).

40. Kleanthous, C. & Walker, D. Immunity proteins: enzyme inhibitors that avoid the active site. *Trends Biochem Sci* **26**, 624-31. (2001).

41. Jakes, K.S. & Zinder, N.D. Highly purified colicin E3 contains immunity protein. *Proc Natl Acad Sci U S A* **71**, 3380-4. (1974).

42. Sidikaro, J. & Nomura, M. E3 immunity substance. A protein from e3-colicinogenic cells that accounts for their immunity to colicin E3. *J Biol Chem* **249**, 445-53. (1974).

43. Wallis, R., Reilly, A., Rowe, A., Moore, G.R., James, R. & Kleanthous, C. (1992). In vivo and in vitro characterization of overproduced colicin E9 immunity protein. *Eur J Biochem*, **207**, 687-95.

44. Wallis, R., Moore, G.R., James, R. & Kleanthous, C. Protein-protein interactions in colicin E9 DNase-immunity protein complexes. 1. Diffusion-controlled association and femtomolar binding for the cognate complex. *Biochemistry* **34**, 13743-50. (1995).

45. Kleanthous, C., Kuhlmann, U.C., Pommer, A.J., Ferguson, N., Radford, S.E., Moore, G.R., *et al.* (1999). Structural and mechanistic basis of immunity toward endonuclease colicins. *Nat Struct Biol*, **6**, 243-52.

46. Kuhlmann, U.C., Kleanthous, C., James, R., Moore, G.R. & Hemmings, A.M. Preliminary X-ray crystallographic analysis of the complex between the DNAase domain of colicin E9 and its cognate immunity protein. *Acta Crystallogr D Biol Crystallogr* **55**, 256-9. (1999).

47. Kuhlmann, U.C., Pommer, A.J., Moore, G.R., James, R. & Kleanthous, C. Specificity in protein-protein interactions: the structural basis for dual recognition in endonuclease colicin-immunity protein complexes. *J Mol Biol* **301**, 1163-78. (2000).

48. Wallis, R., Leung, K.Y., Pommer, A.J., Videler, H., Moore, G.R., James, R., *et al.* (1995). Protein-protein interactions in colicin E9 DNase-immunity protein complexes. 2. Cognate and noncognate interactions that span the millimolar to femtomolar affinity range. *Biochemistry*, **34**, 13751-9.

49. Baggiolini, M., Dewald, B. & Moser, B. Human chemokines: an update. *Annual Review Of Immunology* **15**, 675-705 (1997).

50. Oppenheim, J.J., Zachariae, C.O., Mukaida, N. & Matsushima, K. Properties of the novel proinflammatory supergene intercrine cytokine family. *Annual Review Of Immunology* **9**, 617-48 (1991).

51. Robertson, M.J. Role of chemokines in the biology of natural killer cells. *J Leukoc Biol* **71**, 173-83. (2002).

52. Sun, P.D. & Davies, D.R. The cystine-knot growth-factor superfamily. *Annu Rev Biophys Biomol Struct* **24**, 269-91 (1995).

53. Risau, W. Mechanisms of angiogenesis. *Nature* **386**, 671-4. (1997).

54. Ferrara, N. The role of vascular endothelial growth factor in pathological angiogenesis. *Breast Cancer Res Treat* **36**, 127-37 (1995).

55. Bergmann, D.C. SECuring the perimeter. *Trends Plant Sci* **6**, 235-7. (2001).

56. Leyman, B., Geelen, D. & Blatt, M.R. Localization and control of expression of Nt-Syr1, a tobacco SNARE protein. *Plant J* **24**, 369-81. (2000).

57. Settle, S., Marker, P., Gurley, K., Sinha, A., Thacker, A., Wang, Y., *et al.* (2001). The BMP family member Gdf7 is required for seminal vesicle growth, branching morphogenesis, and cytodifferentiation. *Dev Biol*, **234**, 138-50.

58. Chang, C. & Hemmati-Brivanlou, A. Xenopus GDF6, a new antagonist of noggin and a partner of BMPs. *Development* **126**, 3347-57. (1999).

59. Lee, S.J. & McPherron, A.C. Regulation of myostatin activity and muscle growth. *Proc Natl Acad Sci U S A* **98**, 9306-11. (2001).

60. McPherron, A.C., Lawler, A.M. & Lee, S.J. Regulation of anterior/posterior patterning of the axial skeleton by growth/differentiation factor 11. *Nat Genet* **22**, 260-4. (1999).

61. Ko, T.P., Liao, C.C., Ku, W.Y., Chak, K.F. & Yuan, H.S. The crystal structure of the DNase domain of colicin E7 in complex with its inhibitor Im7 protein. *Structure Fold Des* **7**, 91-102. (1999).

62. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-402.

63. Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. *Numerical Recipes in C*, 503 (Cambridge University Press, Cambridge, 1988).

64. Dascher, C. & Balch, W.E. Mammalian Sly1 regulates syntaxin 5 function in endoplasmic reticulum to Golgi transport. *J Biol Chem* **271**, 15866-9. (1996).

65. Jagadish, M.N., Tellam, J.T., Macaulay, S.L., Gough, K.H., James, D.E. & Ward, C.W. (1997). Novel isoform of syntaxin 1 is expressed in mammalian cells. *Biochem J*, **321**, 151-6.

66. Bock, J.B., Klumperman, J., Davanger, S. & Scheller, R.H. Syntaxin 6 functions in trans-Golgi network vesicle trafficking. *Mol Biol Cell* **8**, 1261-71. (1997).

67. Kosodo, Y., Noda, Y. & Yoda, K. Protein-protein interactions of the yeast Golgi t-SNARE Sed5 protein distinct from its neural plasma membrane cognate syntaxin 1. *Biochem Biophys Res Commun* **250**, 212-6. (1998).

68. Nichols, B.J. & Pelham, H.R. SNAREs and membrane fusion in the Golgi apparatus. *Biochim Biophys Acta* **1404**, 9-31. (1998).

69. Jahn, R. Sec1/Munc18 proteins: mediators of membrane fusion moving to center stage. *Neuron* **27**, 201-4. (2000).

70. Riento, K., Kauppi, M., Keranen, S. & Olkkonen, V.M. Munc18-2, a functional partner of syntaxin 3, controls apical membrane trafficking in epithelial cells. *J Biol Chem* **275**, 13476-83. (2000).

71. Steegmaier, M., Oorschot, V., Klumperman, J. & Scheller, R.H. Syntaxin 17 is abundant in steroidogenic cells and implicated in smooth endoplasmic reticulum membrane dynamics. *Mol Biol Cell* **11**, 2719-31. (2000).

72. Kim, B.Y., Kramer, H., Yamamoto, A., Kominami, E., Kohsaka, S. & Akazawa, C. (2001). Molecular characterization of mammalian homologues of class C Vps proteins that interact with syntaxin-7. *J Biol Chem*, **276**, 29393-402.

73. Dulubova, I., Yamaguchi, T., Wang, Y., Sudhof, T.C. & Rizo, J. Vam3p structure reveals conserved and divergent properties of syntaxins. *Nat Struct Biol* **8**, 258-64. (2001).

74. Assaad, F.F., Huet, Y., Mayer, U. & Jurgens, G. The cytokinesis gene KEULE encodes a Sec1 protein that binds the syntaxin KNOLLE. *J Cell Biol* **152**, 531-43. (2001).

75. Massague, J. & Chen, Y.G. Controlling TGF-beta signaling. *Genes Dev* **14**, 627-44. (2000).

76. IUIS/WHO. Chemokine/chemokine receptor nomenclature. *J Leukoc Biol* **70**, 465-6. (2001).

77. Horuk, R. Chemokine receptors. *Cytokine Growth Factor Rev* **12**, 313-35. (2001).

78. Petrova, T.V., Makinen, T. & Alitalo, K. Signaling via vascular endothelial growth factor receptors. *Exp Cell Res* **253**, 117-30. (1999).

79. Zachary, I. & Gliki, G. Signaling transduction mechanisms mediating biological actions of the vascular endothelial growth factor family. *Cardiovasc Res* **49**, 568-81. (2001).

Conclusion

Further understanding of the evolutionary perspectives of protein-protein interactions can provide valuable information about the structural and functional characteristics of proteins that interact. Based on the hypothesis that proteins that interact must also co-evolve in order to maintain the structurally and functionally relevant features of the binding site, I have presented a method that can quantitatively measure the degree of co-evolution between a family of ligands and a family of receptors by applying a correlation analysis to the phylogenetic distances between the ligands and the phylogenetic distances between the receptors. Using this approach, I showed that chemokines and chemokine receptors have co-evolved. This allows one to make reasonable inferences for identifying potential binding partners for proteins with uncharacterized binding specificity by greatly reducing the search space of possible binding partners to a small subset represented by a region of the protein family's phylogenetic tree. In addition, this method can be easily extended to study other superfamilies as well.

Based on the large amounts of information obtained from genomic efforts, many proteins with uncharacterized function have been discovered. I have applied this co-evolutionary approach to study human CMV-encoded chemokines and chemokine receptors and how these proteins interact with the human immune system. Along with supporting experimental evidence, this analysis suggests the existence of an intricate interplay between the different cytokines, chemokines and chemokine receptors of both viral and host origin. Further examination of the expression kinetics of the CMV-encoded genes in various cellular environments will provide clearer information

regarding the complex interactions between CMV proteins and human proteins involved in the immune response.

The Mlo gene family represents the only sequence-diversified family encoding seven-transmembrane proteins in plants. Mlo is believed to be involved in cell death protection, but its role in this process is unclear. By applying the co-evolutionary analysis to the Mlo protein domains, I showed that there is some evidence for co-evolution of all cytoplasmic loops with the C-terminus. Probable co-evolution between the cytoplasmic domains of Mlo suggests interplay of these domains and interaction with putative partner(s) for Mlo protein function. Although other scenarios are possible, the most likely interpretation is related to a conserved interaction of the cytoplasmic domains with a common binding partner. Future experiments can be done to test the putative co-evolution of the cytoplasmic domains with each other and the C-terminus and to further characterize the function of these domains.

In addition to developing a quantitative measure of co-evolution between two protein families that are known to interact, I extended the co-evolutionary analysis to measure the co-evolution of proteins between the two interacting protein families. By quantitating the co-evolution of proteins, one can begin to make objective inferences of possible candidate binding partners for proteins with uncharacterised binding specificity. I applied this approach to six other protein families and made plausible predictions for interacting partners for orphan proteins in the syntaxin and TGF-β protein families. This approach will become even more powerful as more genes are cloned, thereby filling in the gaps of missing sequence information. While increasing amounts of experimental information can only improve the utility of the co-evolutionary strategy, the co-

evolutionary analysis is a robust and versatile approach to infer which proteins and/or domains are likely to interact when they are parts of larger families that are known to interact.

I have developed and applied a co-evolutionary analysis to protein families that interact. These results have shown the utility of computational methods to functionally characterize the large numbers of newly discovered genes and proteins. As the genomes are nearing completion, the use of computational techniques to gather, process, and synthesize the vast amount of experimental information being accumulated has become increasingly useful. The increased integration of both computational and experimental methods will greatly improve researchers' understanding of basic principles in the biological sciences.

135967

UC San Francisco
LIBRARY
UNIVERSITY OF CALIFORNIA

7132296
3 1378 00713 2296

For reference

Not to be taken from the room.