Title
The value of phylogenetic information in post-genomic protein structure prediction and function discovery

Author
Joachimiak, Marcin Pawel

Publication Date
2002

Peer reviewed|Thesis/dissertation

The Value of Phylogenetic Information in Post-Genomic Protein Structure Prediction and Function Discovery

by
Marcin Pawel Joachimiak

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

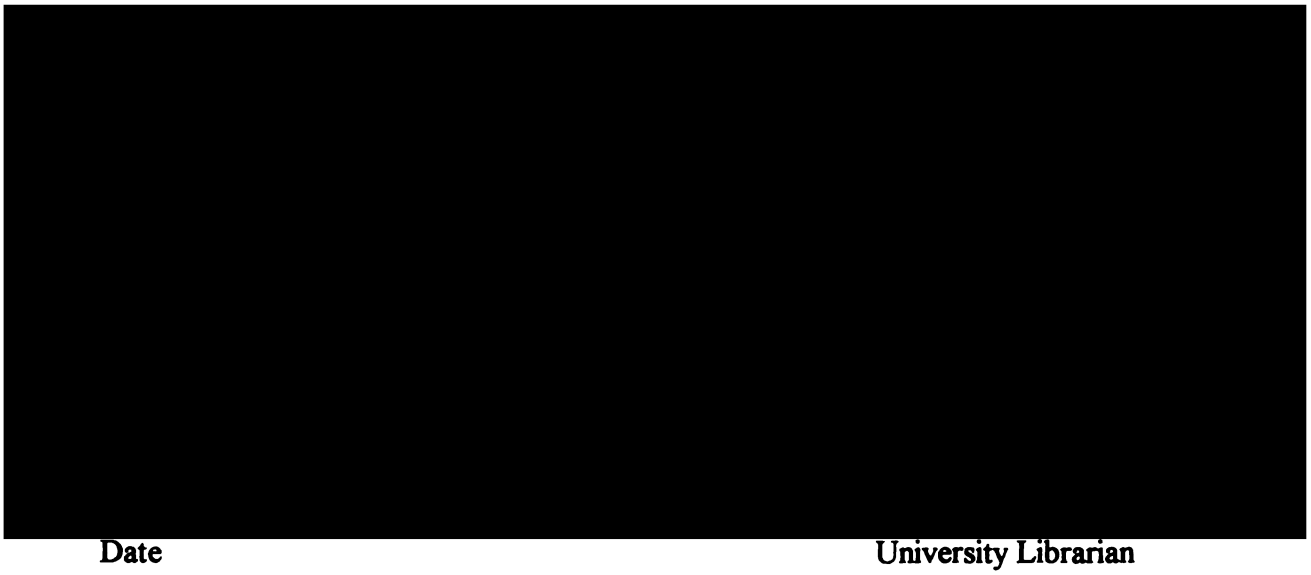
in
Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO



Date University Librarian

Degree Conferred:.....

"You can fool some people sometimes,

but you can't fool all the people all the time."

Bob Marley

Dedicated to

my Mother Grazyna

and

my Father Andrzej

and Ania, my soul mate in life.

Preface

I came to San Francisco and the University of California San Francisco in 1996. I had just completed a B.A. in Mathematics at the University of Chicago. My appetite for integrating mathematical and biological thinking was well honed, with initial experiences of these two very different worlds. While in Chicago, I worked on a recombinant *Triticum aestivum* project with Dr. Robert Haselkorn in the Department of Molecular Genetics and Cell Biology. It was Dr. Haselkorn who provided the initial impetus for my interest in a graduate education, suggesting the UCSF Biophysics program for a young structural biology minded mathematician. The undergraduate experience led me to believe that the biological and the mathematical were quite separate and distant in modern science, with few exceptions. And yet the intrinsic complexities and volumes of modern biological data appeared especially suitable for mathematical analysis, simulation and prediction.

The Biophysics Graduate Program at UCSF was among the academic exceptions where multiple fields were allowed to freely intersect with biology. The Program is represented by excellent faculty and students, all of whom recognized the importance of discussion and imagination to new

approaches grounded in numerical analysis, algorithmic procedures and mathematical constructs. I am greatly thankful to my research advisor Fred E. Cohen M.D. D.Phil. for guidance of my evolution at UCSF. The creative spirit of work and interactions with Fred and his group were something I will never forget. The members of my Preliminary Exam Committee, Dr. David Agard, Dr. Tack Kuntz, Dr. Patsy Babbitt and Henry Bourne M.D., and in particular the latter two faculty who were part of a my Thesis Committee, were also essential to my graduate evolution. It cannot be overstated that these mentors not only recognized the importance of science itself, but also valued the often unintuitive aspects of science, such as communication skills, 'big picture' reasoning, and the experience of graduate education in general.

My thesis would not have been possible if not for many people at UCSF and beyond and I cannot possibly thank them all and enough. A number of former members of the Cohen group are of special importance. Dietlind Gerloff is responsible for my ongoing infatuation with biological evolution and its repercussions for sequence and structure analysis. It was also Dietlind who fully initiated me into the field of protein structure prediction and the CASP contests. Jonathan Blake was my formal introduction to

programming and the JAVA language, skills that are the basis of results in my dissertation. Dirk Walther, Andrew Wallace, Elaine Meng, Xiaohui Du, Paul Harrison and Olivier Lichtarge contributed many crucial suggestions, advice and discussions. I thank Ginger Valen for the warmth and atmosphere in the lab. John-Marc Chandonia, my Cohen group roommate, steered me towards a path of long-term computational and informatic development. Current Cohen group members, Florence Horn, Barney May, Chern-Sing Goh and Anthony Lau, have especially supported me in the final stages of my dissertation. Of my fellow Biophysics Program students, Todd Pray, Kinkead Reiling, Sandy Waugh, Manish Butte, Chuck Sindelar, Andrew Bogan, Peter Chien, Zach Serber, Dan Stone, Alex Schnoes, have been of special significance. However, I am indebted to the positive atmosphere and possibilities offered by the entire UCSF Biophysics program, and especially to Julie Ransom and David Agard. I also sincerely thank Darren Machule and Eric Schnell for changes in scenery. Last but not least there have been significant non-academic components in the road to my graduate degree. My wife Ania has braved the life of a graduate student with infinite patience and understanding. My parents and brother have always encouraged, supported and consoled me. All of these people

positively contributed to my quality of life and experience at UCSF and the city of San Francisco. The people around me have countered the popular belief that the life of a Ph.D. student is characterized by isolation and loneliness.

Dr. Fred E. Cohen was the principal investigator in this work. The data presented in Chapter II has been published in Molecular Medicine.

Joachimiak M.P., Chang C., Rosenthal P.J., Cohen F.E. (2001). The impact of whole genome sequence data on drug discovery--a malaria case study. Mol Med 7(10):698-710.

Abstract

Structure-based drug design of inhibitors against targets with characterized function is the basis for pharmaceutical development worldwide. Meanwhile, genomes of multiple species are providing vast sequence data that is sparsely annotated with structural and functional information. Effective computational methods have become imperative in biological discovery in the post-genomic sequence era.

Structural information provides specific advantages in protein system characterizations and drug discovery. I have developed a method based on primary sequence and phylogenetic information, to identify evolutionary sequence correlations of residue proximal in space. Novel constraints related to conditional probabilities of residue and secondary element contacts are calculated from available structures. I describe an *ab initio* protein contact map prediction method, which relies on iterative application of structural constraints to the evolutionary correlations. The predicted contact maps have features of protein structures not seen in other prediction methods and can be used in combination with other methods to generate protein models.

The post-genomic era provides opportunities to reinterpret data in context of complete catalogs of genes. On the example of a malarial cysteine protease, I have performed protein family analysis, homology modeling and computational screening. A number of antimalarial compounds with low μM activity are identified in cell culture assays. The compounds were selected for predicted binding in unique specificity sites of the malarial protease relative to human protease homologs. Most prescribed antimalarials have significant side effects and lack of similarity between known antimalarials and the compounds identified in this work represents novel routes for antimalarial drug discovery.

The elucidation of protein function in light of whole genome data requires effective computational methods. I have designed and implemented a computer application, JEvTrace, to manipulate and analyze protein family sequences, structures and phylogeny. Based on the ideas of the Evolutionary Trace (Lichtarge, Bourne et al. 1996), a number of algorithmic variations provide new insights into protein function in a family context. Differences and similarities between phylogenetic subclades are analyzed using the primary sequence, tertiary structure, and phylogenetic information. Based on examples of protein

structures with unknown or predicted function, JEvTrace provides evidence for new functions and specificities.

Ab initio prediction of structure and function using primary sequence data directly addresses the problems of the post-genomic sequence era. Biological function and structure and their evolution are incredibly complex and not easily amenable to mathematical and computational representations. Combinations of different methods and data represent an important strategy to account for biological variations and irregularities.

Table of Contents

List of Tables

List of Figures

Prologue

There are many instances of what can be regarded as self-medication in a surprising variety of species (Engel 2002). The specific utilization of plants and minerals by other organisms, in addition to the natural medicine history in our own species, emphasizes the inherent connection between the ability to survive and awareness of biochemical processes. For example, a fundamental mechanism is the phenomenon of thirst which at the molecular level can be regarded as control of osmolarity (Stricker and Sved 2000). With rapidly expanding molecular biology knowledge, detailed connections between the macro scale of organisms and the nano molecular scale of biological phenomenon are increasingly feasible.

A biological revolution occurred in the second half of the 20th and the beginning of the 21st century. The discovery of protein structure dates to the determination of the first protein crystal structure, that of human hemoglobin (Muirhead et al 1967). This discovery introduced a new world of macromolecules to biology, chemistry and physics. The DNA and RNA polyribonucleic acids were macromolecular structure precursors to proteins in the chronology of

biomolecular structure determination (Watson & Crick 1953). Although genetics, evolution and significant functional speciations are understood to occur at the nucleotide level of life, it is the vast implications of protein structure and function for health, disease, industry and agriculture that attracted the most interest. Structural biology has provided a basis for the molecular paradigm in modern biology. Structural data on protein function and structure has led to unraveling of the molecular aspects of human health and disease. These aspects provide knowledge for improved health and human capacity as well as direct routes to drug discovery efforts.

Many natural medicine remedies are practiced worldwide, some with millennial histories of use by human cultures (Swerdlow and Johnson 2000). Discoveries made by our ancestors have been revisited at the molecular level by modern biomolecular scientists. For example salicylic acid (meaning "acid from willow tree"), commonly known as aspirin, may not have been approved for use by current drug screening standards due to its side effects profile. Interestingly, the hemorrhagic side effects of salicylic acid seem to be mitigated by a second ultra low dose of the same molecule (Aguejouf, Malfatti et al. 2000). Among the features of salicylic acid is the covalent nature of the

drug interaction with prostaglandin H2 synthase (Loll, Picot et al. 1995). Molecular complementarity as first described by L. Pauling (Pauling 1974), also referred to as molecular specificity, plays a prominent role in the drug efficacy landscape. One of the reasons for the problematic side effect profiles of covalent inhibitors as drugs is there ability to inhibit homologous proteins. It has been argued that mixtures of compounds in natural medicines, representing a nonlinear gain in pharmacological activity, can diminish the side effects of principle pharmacological activities (Swerdlow and Johnson 2000).

The complex activity of pharmacological mixtures is one of the reasons for slow progress in molecular dissection of natural product activities (Swerdlow and Johnson 2000). The dried seeds of *Coffea arabica*, the leaves of the *Thea* family and the seeds of the *Theobroma Cacao* tree provide prominent examples of molecular activities stemming from hundreds if not thousands of compounds. Caffeine is among the few closely studied molecules from the mixtures present in coffee and tea. However, caffeine alone does not fully account for the total activity of the world's favorite hot drinks. The active mixture interpretation applies to nearly everything in our diet. Individual species have distinct metabolic and

small molecule profiles, dependent on genetic and environmental factors. The activity of small molecule mixtures against a range of drug targets is a poignant example of the complexity in the molecular basis of life.

The biological implications of protein structure were recognized long before the advent of protein structure determination. In 1941, W.T. Astbury et al described proteins as "essentially a pattern of charges associated with a close-packed forest of side chains, the polypeptide chain being an infinitely variable device for building up distributions of charges and presenting them to given chemical environments." (Astbury, S. et al. 1941). A structural basis for biological function formed the impetus for protein structure modeling and experimental determination. The beginnings of heuristic approaches to assigning, modeling and predicting protein structure date back to the early days of x-ray crystal structure determination of compounds related to proteins. In early structural studies, structures of protein related compounds were coupled with the development of protein structure heuristics related to molecular constraints of the polypeptide backbone in light of lattice theory and diffraction data (Huggins 1943). Regular protein secondary structures formed by networks of repetitive hydrogen bonds

were postulated in 1951 by Linus Pauling et al (Pauling, Corey et al. 1951). These structures correspond to the alpha helix and the parallel and antiparallel beta-pleated sheets.

The use of protein structure models to test hypothesis and subsequent development of heuristics is not unrelated to practices in organic chemistry. Indeed the activity of building models under abstract constraints can be recognized as an ancient one. Philosophically, heuristics in modeling and prediction can often be reduced to the presence of significant homology of two entities with respect to a single property, hence suggesting similarity in other correlated properties.

As of February 5th 2002 there are 17428 publicly available protein structures in the Protein Data Bank (Bernstein, Koetzle et al. 1978). Proteins are now known to harbor specific structural preferences for amino acids in alpha helices, even with respect to the chain direction (Blaber, Zhang et al. 1993). Similar data is available for beta-sheets (Otzen and Fersht 1995). Sequence to structure correlations in local (Bystroff, Simons et al. 1996; Bystroff, Thorsson et al. 2000) and global protein structure (Grishin and Phillips 1994; Gromiha and Selvaraj 2001) are providing important information on the sequential

aspects of protein folding and structure. And a large list of protein structural motifs (Han and Baker 1995; Han and Baker 1996; Han, Bystroff et al. 1997) has been derived from biological databases. Sequence to structure correlations include motifs correlated with specific functions such as zinc fingers and RNA binding.

In spite of considerable progress, a number of important computational biology problems remain unsolved. In fact, a large number are not yet attempted due to ongoing accumulations of genome-wide data sets. The implications of knowledge of protein structure and function are clear. Human health and physical capacity have new potentials based on our abilities to measure, model and alter biological function. Industrial and agricultural applications are just beginning to be explored. It is exceedingly important to include history in current bodies of knowledge. The process of evolution of life is one of the most meaningful and fascinating histories available.

Protein evolution in the form of diversity and plasticity contains information on the structure and function of proteins. The following three chapters demonstrate directed attempts to advance post-genomic computational biology in the realms of evolutionary correlations and protein structure protein, drug design in

a protein family context, and computational methods for protein function discovery based on the ideas of the Evolutionary Trace (Lichtarge, Bourne et al. 1996).

Chapter I

Protein structure prediction by combining cooperative sequence evolution at spatially proximal sites with structural constraints

Introduction

A Structural Introduction to Protein Structure Prediction

The astounding complexities and resulting properties of protein structure arise from a twenty amino acid alphabet. Even more intriguing, the amino acid sequence of a protein contains the necessary information for forming a folded, functional molecule (Anfinsen, Haber et al. 1961; Epstain, Goldberger et al. 1963; Anfinsen 1973). This property of proteins has been termed 'Anfinsen's dogma'. The associated Levinthal 'paradox' of protein folding (Zwanzig, Szabo et al. 1992), states that a protein would require 10^{N-14} seconds (where N is the length of the protein) to sample all possible conformations of the polypeptide chain. This result is based on thermodynamic principles and the assumption of 10^{14} conformations changes per second (based on fast spectroscopy measurements). For a 40 residue protein this would take 10^{16} seconds. This clearly does not occur in nature. The dogma and the paradox have implied that there are general rules governing the formation of protein tertiary structure. Existence of physical rules has

suggested feasibility of algorithmic approaches to predicting tertiary structure from sequence.

Protein structural requirements of amino acid side chains include features of amino acid chemistry such as molecular volume, electrostatics, hydrogen bonding potential and hydrophobicity. Less understood are features endowed upon amino acids by protein tertiary structure such as buried surface area, quantity and quality of neighboring residues, dihedral angle preferences and chirality constraints. A long recognized entropic constraint on protein structure has been local backbone information with specific periodicities corresponding to the alpha helix (i,i+4) (Pauling and Corey 1951) and beta strands (i,i+2 & i,i+6) (Pauling and Corey 1951). More intricate periodicities in tertiary interactions such as the i+4+n periodicity in the packing of adjacent alpha helices separated by a loop, and i+2+n for adjacent beta strand packing, have previously been applied to protein tertiary structure prediction (Cohen, Sternberg et al. 1982). Recently structure-based heuristics for the more complex phenomenon of register between secondary elements have been developed, for example in antiparallel beta sheets (Hutchinson, Sessions et al. 1998).

Structurally, the amino acids of a polypeptide backbone partition to the surface, the hydrophobic core, and a boundary layer partially accessible to solvent, such as active site clefts (Lichtarge, Bourne et al. 1996). Conformational states that result in residue solvent accessibility changes complicate this interpretation. Nevertheless, the sequence patterns indicative of residue burial or exposure are among the strongest signals in multiple sequence data (Benner, Badcoe et al. 1994; Gerloff, Joachimiak et al. 1998; Gerloff, Cannarozzi et al. 1999).

It has been known for some time that structure is more conserved than sequence (Chothia and Lesk 1986; Chothia and Lesk 1987; Flores, Orengo et al. 1993). It follows that in protein evolution, preservation of structure takes precedence over conservation of sequence. Detailed analysis of structural homolog comparisons has lead to the conclusion that a relatively small (on the order of 30%) fraction of residues are required to define a common hydrophobic core, and hence a common fold (Russell and Barton 1994). The remaining fraction of residues can adopt unique conformations, evolving beyond recognition of structural similarity - such structural variations can occur even with conservation of function (Russell and

Barton 1994). The conservation of structure has important consequences for protein structure prediction and evolution.

It has been demonstrated that similarity measures based on protein residues interactions cluster the known protein structures more strongly than sequence comparisons (Godzik, Skolnick et al. 1993). In addition to the conservation of structure, it seems that this result has two underlying reasons. Firstly, there is a bias in experimental studies based on purification, stability and crystallization of proteins. This bias is represented in any structural data set. Secondly, the information content of pairwise amino acid contacts is less than the information present in multiple sequence alignment (MSA) data. Multiple homologous sequences represent an ensemble of information that is reduced to a unified sequence representation when embedded in a two dimensional matrix encoding protein contacts. Such contact maps are to protein structure as amino acid properties (reduced alphabets) are to multiple sequence data. It is interesting to speculate that clustering structures based on chemical properties such as charge and volume, would lead to a tighter clustering than that based on the twenty amino acids.

A feature of protein residue interactions is sequence separation, defined by the number of residues in sequence between a pair of residues. Short range interactions are strongly influenced by regular secondary structure protein backbone constraints (Pauling and Corey 1951; Pauling, Corey et al. 1951). Constraints of these local interactions are also evident in the Ramachandran map of preferred residue conformations (Edsall, Flory et al. 1966) and in the occupied conformational space of protein sidechains (Walther and Cohen 1999). Medium range interactions are characterized by a sequence separation of at least 3.6 residues (the periodicity of the alpha helix) and less than the length of the shortest secondary element in the structure. Long range interactions are usually understood to be residue-residue contacts with a sequence separation of more than the shortest secondary structure element. Long range interactions are the multiple interactions that constitute tertiary protein structure in the form of contacts between secondary structure elements.

The accumulation of evidence for the evolutionary significance of residue contacts involved in protein folding, the folding nucleus and global determinants of protein stability is steadily increasing (Shakhnovich, Abkevich et al. 1996; Dosztanyi, Fiser et al. 1997;

Michnick and Shakhnovich 1998; Ptitsyn 1998; Ptitsyn and Ting 1999; Mirny and Shakhnovich 2001). Recently Gromiha & Selvaraj (Gromiha and Selvaraj 2001) confirmed the correlation between protein contact order (Plaxco, Simons et al. 1998), a measure of the sequence separation between residue contacts in a protein, and the folding rate of two-state proteins. The implication is that the larger the number of long range interactions in the native state, the more pronounced the thermodynamic barrier between the disordered and native states. In the case of beta sheets, structures with a high contact order, medium and long range interactions represent the basis of beta sheet formation (Pauling and Corey 1951).

Protein structure is the ultimate source of data on residue-residue interactions as well as constraints determining specific sequence to structure correlations. The PDB database (Bernstein, Koetzle et al. 1978) of known protein structures exists in manually curated (SCOP (Lo Conte, Brenner et al. 2002)) and automatically generated (CATH (Orengo, Bray et al. 2002)) classifications. These databases organize protein structures into families, folds and superfolds based on evolutionary and supra-structural homology interpretations. The significant volume of available protein tertiary structure data provides

structural knowledge for developing, testing and applying protein structure prediction methods.

Proteins from an Amino Acid Interaction Perspective

The twenty natural L-amino acids, including the amides and carbonyls of the protein backbone and the charged N and C termini, are the chemical units of protein interactions. These chemical units are associated with alphabets of interactions, for example pairs of interacting residues. Considering pairs of residues, there are 190 possible unique pairwise interactions assuming directionality of the polypeptide chain. The directionality assumption implies that the interactions of X with Y and Y with X are not symmetric. This is in fact the case ((Sippl 1990) and this work).

Many protein intra-residue interactions, such as contacts between secondary structures, involve more than pairs of residues. A combinatorial explosion in possibilities accompanies higher orders of residue interaction: 1140 unique amino acid triplets, 4845 quadruplets etc. In general, higher order or multi-body computational problems are confronted with physical and

algorithmic limitations. These problems are called NP-complete or NP-hard, due to their non-polynomial solutions. Due to these circumstances, pairs of resides have become an accepted simplification in modeling proteins and their interactions. *Ab initio* protein structure prediction can be viewed as the prediction of residue-residue distance constraints. Experimental protein structure determination consists of measuring atomic distances or distance constraints based on protein crystal x-ray diffraction or NOE coupling data. High quality NMR structures usually have on the order of 20 or more constraints per residue (Holmbeck, Foster et al. 1998; Muskett, Frenkiel et al. 1998; Morshauser, Hu et al. 1999). Low resolution protein models have been determined based on sparse distance restraints (Lund, Hansen et al. 1996; Chelvanayagam, Knecht et al. 1998), down to one or two constraints per residue (Aszodi, Gradwell et al. 1995). It has also been shown that protein structure can be determined even with ambiguous distance restraints (Nilges 1995; Nilges 1997; Nilges, Macias et al. 1997). From the perspective of the protein universe, a small number of stringent distance constraints are the most specific in recognizing specific protein folds (M.M. Young and I.D. Kuntz, personal communication). Therefore the reduction of the multi-body space of protein

residue interactions to pairs (n = 2) is acceptable, especially for the purposes of protein structure prediction.

There are a number of approaches for defining contacts between residues in protein monomers. Due to computation and memory limits, information is sacrificed either in the ability to sample data or in the resolution of the computational model. A two dimensional matrix can be used to represent protein residue-residue interactions. Based on Cα-Cα distance calculations in known protein structures, and a fixed distance cutoff included in the residue-residue contact definition, binary contact information is stored in an *n* by *n* matrix (where *n* is the sequence length). This method has been extensively developed, with broad applicability to sequence homology searches, fold recognition and sequence analysis.

Available protein structures have been used in a knowledge-based approach to determine pairwise energies of residue-residue interactions (Crippen and Viswanadhan 1984; Miyazawa and Jernigan 1985). Attempts to define and search protein families based on pairwise contact preferences (Miyazawa and Jernigan 1993; Rodionov and Johnson 1994) led to the development of methods based on residue-residue interaction matrices. Miyazawa and Jernigan (Miyazawa and

Jernigan 1993) introduced the assumption that large samples of amino acid substitution data could approximate the actual interaction energies of residues in protein structures. The inverse folding problem (Blundell 1991; Yue and Dill 1992) culminated with sequence to structure threading (Bowie, Luthy et al. 1991; Godzik, Kolinski et al. 1992), a method exploiting dynamic programming to assess the preference of a sequence for candidate structures. More complex potentials relying on calculations of mean force and distance distributions of protein residue-residue (Sippl 1990; Maiorov and Crippen 1992) are noted improvements. Another instance of pairwise interactions is seen in lattice models for protein folding and thermodynamic calculations (Yue, Fiebig et al. 1995; Harrison, Chan et al. 1999; Harrison, Chan et al. 2001).

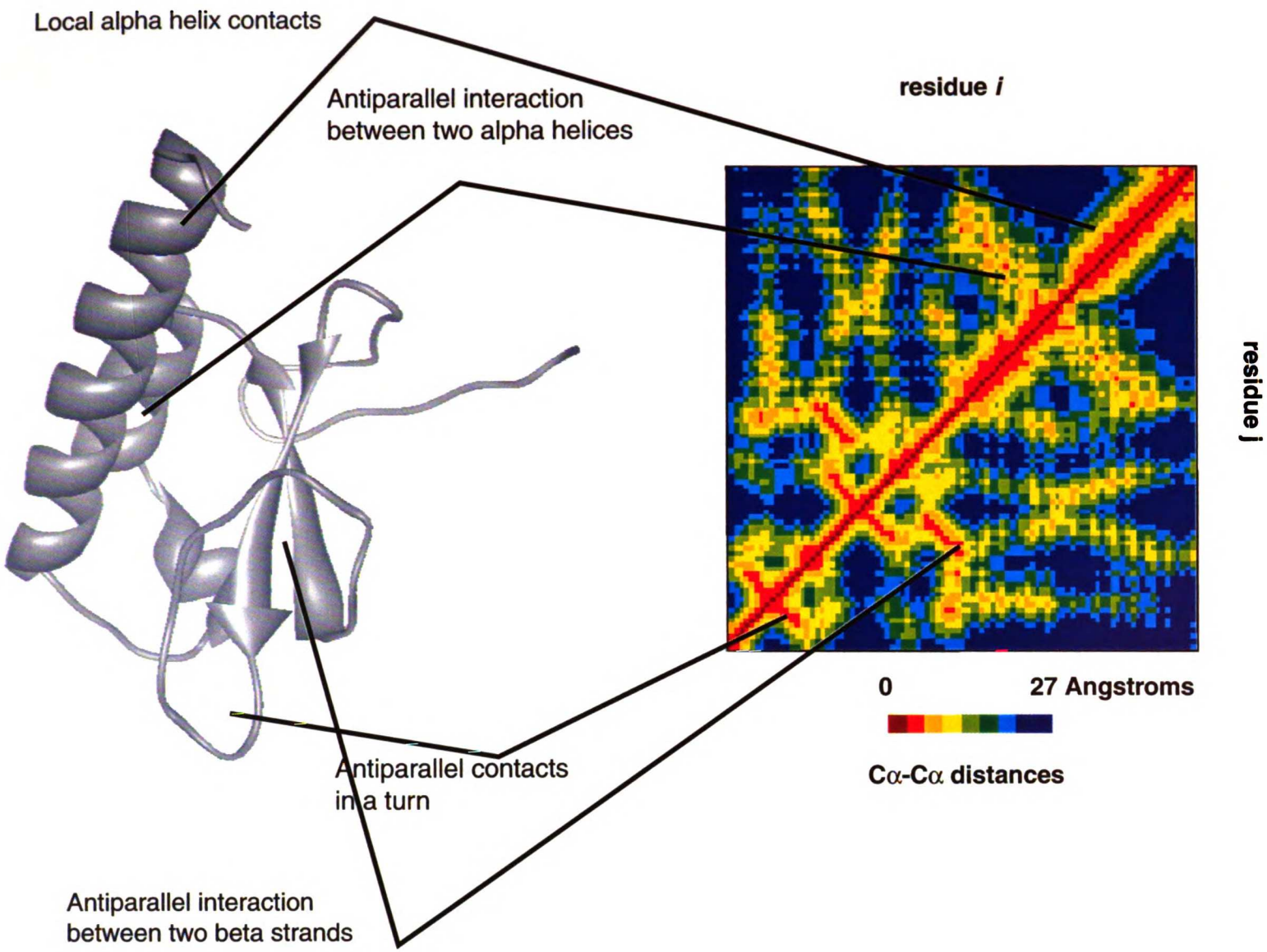
With the assumption of a common structure, sequence differences represent multiple sequence solutions to protein folding. M. Sippl has introduced the notion of structural ensembles within a family of sequences with common structure (Sippl 1990). This definition allows interpretation of protein sequence structure and data with additional parameters. The parameters include asymmetry of pairwise residue interactions (assuming an N to C terminus chain direction) and approaches to normalizing pairwise

interaction frequencies. The effect is an increase in information content and more detailed estimates of pairwise protein residue interaction preferences. A related ensemble approach has been applied in the prediction of consensus protein structure from a protein family ensemble (Gobel, Sander et al. 1994).

Two dimensional protein contact maps represent a simplification of protein interactions. This reduction allows additional levels of analysis, such as global features of protein contacts. Features of protein contact maps include local secondary structure patterns of alpha helix and beta-sheet at specific periodicities of i+3.6 and i+2. Clusters of contacts with specific orientations relative to the diagonal of the contact map correspond to contacts between secondary structures in parallel or anti-parallel orientations (Figure 1). Non-local contacts between secondary structures can also exhibit contact periodicities related to the interacting faces of alpha helices and beta-sheets (Figure 1). Common irregularities in protein structure lead to scattered clusters of contacts with few if any periodic or spatial trends (Figure 1). Contact maps also capture information on the orientation and chirality of supra-secondary structure arrangements in protein tertiary structure.

Figure 1

An example of patterns in protein residue contacts represented by a two dimensional residue-residue contact map. On the left is a structure of the YlxR protein from *Streptococcus pneumoniae* (PDB: 1G2R). The ribbon representation was generated with Chimera (Huang 1996). On the right is a two dimensional matrix of residue-residue distances. The contact matrix graphic was generated with WEBMOL (Walther 1997). The correspondence of regular patterns of secondary structure element interactions to features in the contact matrix is shown.


Local alpha helix contacts
Antiparallel interaction
between two alpha helices
residue i
residue j
0
27 Angstroms
Cα-Cα distances
Antiparallel contacts
in a turn
Antiparallel interaction
between two beta strands

Higher orders of residue interaction are accompanied by a nonlinear increase in the potential interaction space, with the number of interactions proportional to n^k (where n is the sequence length, k is the order of interaction). However, sequence evolution demonstrates that the conservation of residues within protein families is sequentially and spatially specific (Garvin and Hardies 1991; Lichtarge, Bourne et al. 1996; Mirny and Shakhnovich 2001). Therefore, interpretations that allow for meaningful traversal of the interaction space across the possible orders of interaction are especially desirable.

Evolutionary Residue-Residue Correlations

The theory of neutral drift in protein evolution encompasses random sequence changes with neutral effects on protein structure and function (Kimura 1968). These changes are the hallmarks of conservative amino acid mutations and protein structural plasticity. In addition to neutral drift, proteins undergo evolutionary selection for specific properties. Sequence differences coinciding with protein family speciation events reflect the underlying structural and functional requirements of specific in vivo roles.

Even though it has been shown that there exist sequence evolution trends in conservation of hydrogen bonding or volume, these trends are not consistent across evolutionary time scales and protein families (Chelvanayagam, Eggenschwiler et al. 1997). It appears that the standard parameters are insufficient to fully understand conservation patterns in a protein family. Our limited understanding of protein evolution forms a barrier to making general connections between sequence, structure and function. Since sequence data is most prominent in the literature and databases, followed by function and structure, sequence based structure and function predictions methods are of clear value.

Based on examples of protein function discovery, it appears that the evolutionary information represented by the structure of subclades in a phylogeny (Lichtarge, Bourne et al. 1996; Mirny and Shakhnovich 2001), is a source of information across different orders of residue-residue interactions. Surface epitopes and partially buried clefts form the basis of protein function due to their interactions with solvent. These structural features are formed by sets of residues restricted in space. Correlations of sequence to structure have been observed in compensatory mutations (Poteete, Sun et al. 1991; Baldwin,

Xu et al. 1996), protein structural plasticity (Baldwin, Hajiseyedjavadi et al. 1993; Gerstein, Sonnhammer et al. 1994; Vetter, Baase et al. 1996; Atwell, Ultsch et al. 1997; Taverna and Goldstein 2002) and experimental sequence to structure correlation analysis, such as site-directed mutagenesis and alanine scanning (Wells 1991).

From an evolutionary perspective, amino acids that are conserved across species tend to signify functional sites or structural determinants of protein families. These functional and structural categories can have considerable overlap, especially in context of correlated sequence changes. Protein surface epitopes have defined volumes and orientations, features determined by neighboring (potentially non-functional) residues and contributions of more distal regions in a structure. Therefore, models of protein structure evolution should not separate structure from function.

Heuristics based on evolutionary sequence data have been developed for secondary structure prediction (Rost and Sander 1994; Thompson and Goldstein 1997) and threading (Defay and Cohen 1996; Jones, Tress et al. 1999; Panchenko, Marchler-Bauer et al. 1999). However, current understanding of protein evolution, chemistry, and dynamics is still limited. Due to a lack of general principles, there is

little predictive value for sequences without representative structures or no identifiable homologues, as seen in the pressing problem of genome annotation and function prediction (Chapter III). A similar limitation exists in the field of protein structure prediction. However, due to the complexity of proteins structure, structure prediction inherently requires more variables, heuristics, and input data relative to the problem of elucidating function. The evidence for this is the relative success of sequence based function prediction relative to sequence based structure prediction methods.

Determinants of protein sequence evolution are not without effect on protein structure. In particular, specific pairs of interactions undergo evolutionary selection for sequence constraints such as codon bias or amino acid availability. Nucleotide codon bias has measurable influence on protein sequence evolution across single genomes (Singer and Hickey 2000). Amino acids occur with variable abundance in nature and ecological and biochemical constraints must contribute to their relative frequencies in different species. As observed in the sequence databases, the occurrence of individual amino acids in proteins is based on complex factors ranging from influences of the genetic code on amino acid selection, to

residue interaction preferences and constraints on local and global protein structure. These factors are difficult to account for when constructing sequence alignments and phylogenies, and present a limitation for methods relying on phylogenetic data. Nevertheless, percent sequence identity alone approximates the functional speciations within a protein family to a degree that is useful for computational biology predictions.

A method has been developed to identify residues invariant during the evolution of a protein family (Lichtarge et al JMB 1996). Clusters of these residues on a protein surface strongly correlate with functional sites (Chapter II and Chapter III). It is also known that the sequence conservation of oligomeric enzyme subunit interfaces is correlated with overall sequence conservation (Grishin and Phillips 1994), and that correlated mutations contain information about protein-protein interactions (Pazos, Helmer-Citterich et al. 1997). The information present in the branching pattern of subclades in a phylogeny has been applied to the prediction of protein-protein binding with some success (Goh, Bogan et al. 2000; Johnson and Church 2000; Pazos and Valencia 2001) as well as function discovery from sequence and structure data (Chapter III).

Sequence correlations based on either pure MSA data or phylogenetic information, have been applied to fold recognition (Olmea, Rost et al. 1999), to the determination of relationships between structure and sequence patterns (Selbig and Argos 1998), and to the prediction of residue contacts based on a neural network (Fariselli and Casadio 1999; Fariselli, Olmea et al. 2001). The latter neural network method claims the highest accuracy and improvements over random models, with 25% and 8 fold respectively. On average the method was only 16% accurate, which was insufficient to generate useful distance constraints.

Multiple definitions of sequence correlation and methods identifying these types of sequence changes have been developed. The implementations range from pattern based methods to calculated vectors of amino acid properties (Altschuh, Lesk et al. 1987; Altschuh, Vernet et al. 1988; Korber, Farber et al. 1993; Gobel, Sander et al. 1994; Neher 1994; Shindyalov, Kolchanov et al. 1994; Singer, Oliveira et al. 1995; Thomas, Casari et al. 1996; Chelvanayagam, Eggenschwiler et al. 1997; Pollock and Taylor 1997; Fariselli and Casadio 1999; Pollock, Taylor et al. 1999; Fariselli, Olmea et al. 2001). Only a few of these methods attempt to use phylogeny structure information explicitly (Shindyalov, Kolchanov et al. 1994;

Chelvanayagam, Eggenschwiler et al. 1997; Pollock, Taylor et al. 1999). Overall, the individual performance and applicability of these methods for practical purposes of protein structure prediction remain unsatisfactory (Orengo, Bray et al. 1999), especially considering drug design and function discovery needs.

Many sequence correlation methods use MSA or evolutionary data converted into scalar, vector or matrix representations of properties or correlations of properties between MSA positions. Patterns of speciation events that result in distinct phylogenetic subclades are a feature of protein sequence evolution. These subclades can be interpreted as an ensemble of related protein sequences and structures with expected similarity in function and sequence correlations. Such a model of sequence family evolution effectively increases the information content per MSA position relative to numerical constructs used to analyze MSA data alone.

Residues involved in binding are not the only residues conserved during evolution. The dynamic aspects of protein function such as conformational changes (Gerstein, Lesk et al. 1994) and thermal motions (Hoh 1998) are also significant evolutionary constraints. It can be expected that the various constraints on protein sequence, structure

and function in the context of evolution will also be determinants of structure and function. In addition to structural and genetic constraints, functional residues are constrained by selection for an *in vivo* role. Protein function is often characterized by multiple surface and binding partners (dos Remedios and Thomas 2001; Xiong, Stehle et al. 2001), and not only in the case of membrane or structural proteins (Lichtarge, Bourne et al. 1996). There are less recognized attributes of function, such as cellular localization, signal and tag sequences, and sites of post-translational modification, with distinct evolutionary pressures. For example, protein surfaces have undergone differential selection dependent on their subcellular localization (Andrade, O'Donoghue et al. 1998). However, only with sufficient data is it possible to survey the effects of a protein feature on protein evolution. In the case of subcellular localization, useful in predicting the amino acid composition of protein surfaces, there is sparse data on the localization of individual sequences.

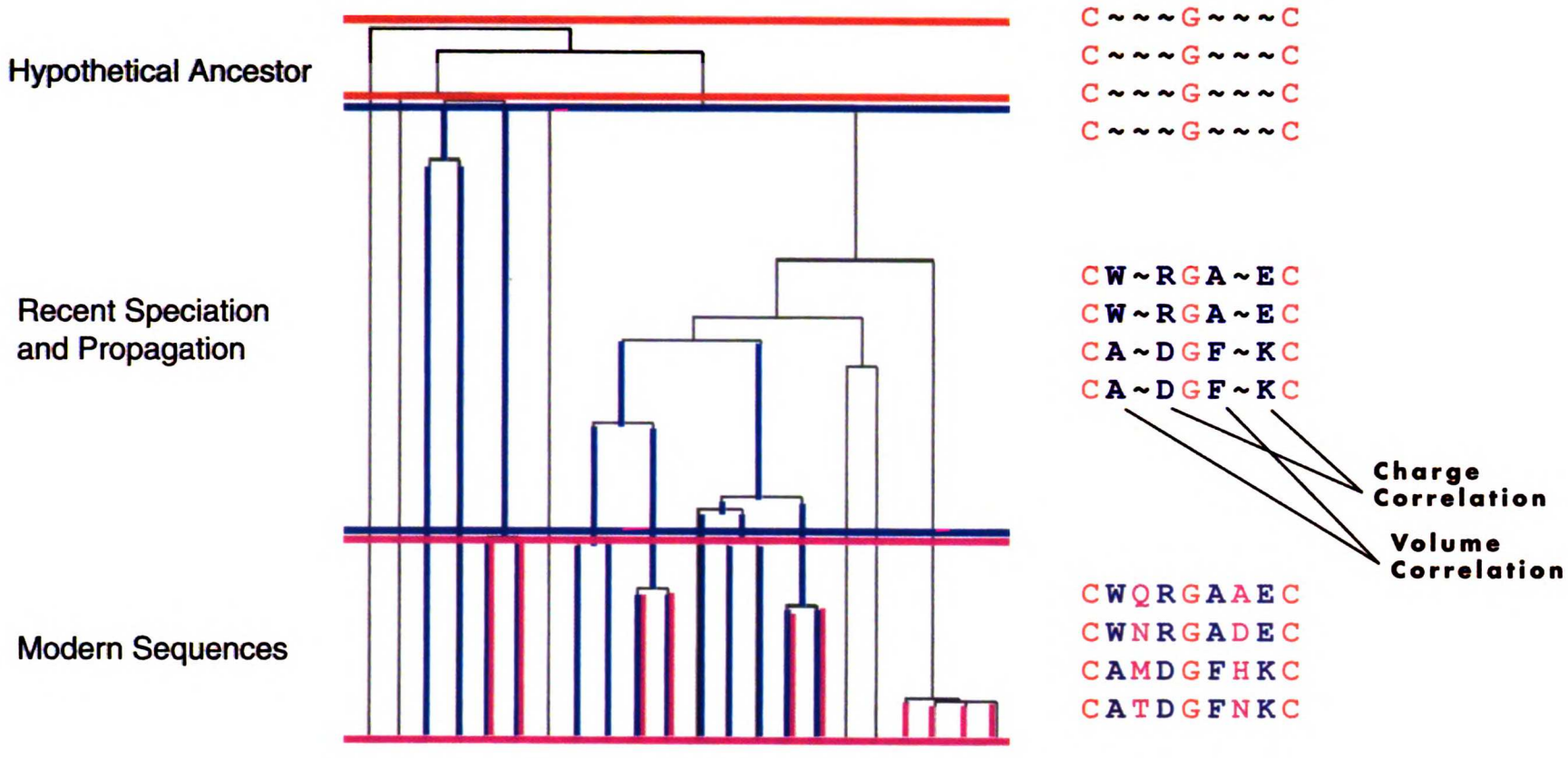
There are serious difficulties with dissecting true evolutionary correlations of residues from correlations arising in the random background of mutation and inheritance (Pollock and Taylor 1997). Pollock & Taylor point out that increasing the number of taxa in the

phylogeny decreases the random background. For methods utilizing phylogenetic data this has the statistical effect of increasing the likelihood of false positive correlations. The current method redefines subclade invariant residues, as originally defined in the ET method (Lichtarge, Bourne et al. 1996), in context of protein contact map prediction. The survey and test sets for predicting pairs of amino acid contacts from sequence and phylogenetic data are combinations of representative protein structure, MSA and phylogenetic information (see Chapter III). The method has the potential to include residue interactions of higher order than the standard pairs of residues.

An idealized example of evolutionary correlation in a phylogeny is shown in Figure 2. The phylogeny is divided into evolutionary timescales, corresponding to eras of mutational events. The dendogram of evolutionary relationships can be interpreted as sets of residues conserved in specific subclades of the phylogeny. The subclades of the phylogeny closest to the hypothetical ancestor represent the positions with the most stringent constraints during evolutionary selection. Given sufficient sequence data, the nodes in the phylogeny closest to individual sequences correspond to mutational events that

Figure 2

A schematic of evolutionary time scales in a phylogeny and their coincidence with specific variability patterns in the multiple sequence data. For each colored time scale, the invariant positions in the multiple sequence data are shown in a corresponding color. Two types of amino acid evolutionary correlation are the conservation of charge and volume properties. Since most modern sequence events correspond to neutral drift, while subclades deep in the phylogeny are difficult to reconstruct for lack of data, evolutionary correlations are expected to occur in the middle time scales of reconstructed phylogenies.


Hypothetical Ancestor
Recent Speciation and Propagation
Modern Sequences
C~~~G~~~C
C~~~G~~~C
C~~~G~~~C
C~~~G~~~C
CW~RGA~EC
CW~RGA~EC
CA~DGF~KC
CA~DGF~KC
Charge Correlation
Volume Correlation
CWQRGAAEC
CWNRGADEC
CAMDGFHKC
CATDGFNKC

are either random changes in single sequences or unique determinants of functional properties not found in other members of the family. The multiple branches of the phylogeny contained between the early and late regions of the evolutionary timescales represent sequence selection events related to aspects of folding, stability and function. Random drift also contributes to evolutionary correlations - random mutations that are compensated for by mutations elsewhere in the protein, will be inherited as a correlated pattern. Random mutations with a neutral effect on the protein's fitness are more likely to undergo further random mutation, and will present different sequence variability patterns. Zuckerkandl and Pauling described living systems as being "constantly abolished and simultaneously preserved", using an analogy from Hegel (Zuckerkandl and Pauling 1965). It is the middle era of the evolutionary time scale that represents evolutionary palimpsests, the writing of new material over older material, of historical speciation and random drift events. Modern subclades are accompanied by increasing levels of neutral drift.

The subclade branching pattern seen in all phylogenies reflects the biologically universal process of evolutionary divergence. Adaptation to new requirements and compensation

for undesirable changes are the dominant forces in evolution. Undesirable changes are random and requirements for structure and function are not uniform across species. These two conditions lead to unique sequence changes in different members of a phylogeny. Inheritance and propagation of speciation events results in further divergence and subsequent appearance of new subclades within the phylogeny.

Phylogenetic information is orthogonal to primary sequence and structural information, serving as an additional dimension in which to understand the underlying correlations. The goals of this evolutionary sequence correlation approach, aside from an attempt to predict protein structure, were to understand protein structure in context of residue-residue contact features, and to merge the concept of an evolutionary sequence mutation space with the realm of tertiary protein structure. Insights provided by analysis of protein evolution in the context of protein structure add to the growing body of knowledge on the subtle correlations relating biological sequence, structure and function.

Results

Formulation of the pairwise correlation model, structure prediction heuristics and construction of the prediction algorithm was preceded by a survey of correlated amino acid contacts in a set of representative protein families. The initial definition of residue evolutionary correlation consisted of an invariant position pair in partitions of the phylogenetic tree (Lichtarge, Bourne et al. 1996). The intervals were defined by the sequence identity of subclades within the partitions and spanned from the hypothetical ancestor of the phylogeny to the individual modern sequences. For a definition of residue contacts see Methods. The serine proteases are chosen as a representative example, but other surveyed families included the HIV proteases, creatine kinases, lactalbumins and lysozymes, amino-aspartyl transferases, and the barnases. Choice of protein family was limited by sequence data and evolutionary diversity, as well as high resolution crystal structures of representative family members.

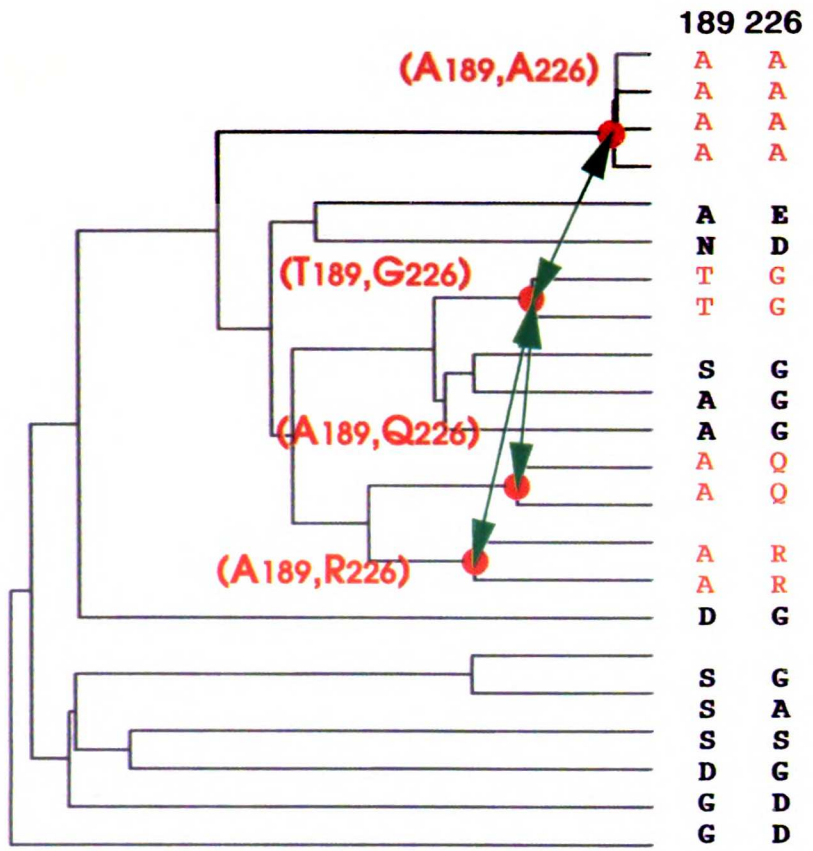
The serine proteases are a protein family with the feature of an extended binding site and rigid structural constraints related to the binding site volume. A number of

correlated residues were in the vicinity of the conserved catalytic triad and in distal portion of the binding site (Figure 3). Detailed phylogenetic analysis indicated that the nature of the evolutionary correlation was layered, with certain subclades representing a change at one site, other subclades at another site, and some at both sites (Figure 3). Subclades with a 'double mutation' relative to other subclades were bridged by a subclade sharing only a single change with the 'double mutant' (Figure 3). Noticeably, a number of seemingly reasonable correlations were in fact distal in space, representing the false positive component of the evolutionary correlation data. The results of the test survey of protein family evolutionary correlations confirmed the presence of pairwise residue contact correlations in a range of protein families.

False positive contact correlations including the background noise in protein sequence evolution were first addressed in the model of evolutionary correlation (Figure 3). The MSA data used to derive the correlations was successively filtered position by position using criteria of sequence variability, amino acid entropy (defined as the number of rotatable bonds per residue), and number of gaps (see Methods). These measures, aside from permitting

Figure 3

An example of protein family correlated contact survey results for a subset of the serine protease family. The residues conserved in specific subclades are shown on the phylogeny (left). The corresponding columns from the MSA highlight the correlated sequence changes. The structure graphic (right) highlights the proximity of the correlated positions in space in a crystal structure of the chymotrypsin member of the family. In this case, the distance between the Cβ of serine and Cα of glycine was 6.5 Å. The catalytic triad is show in blue.



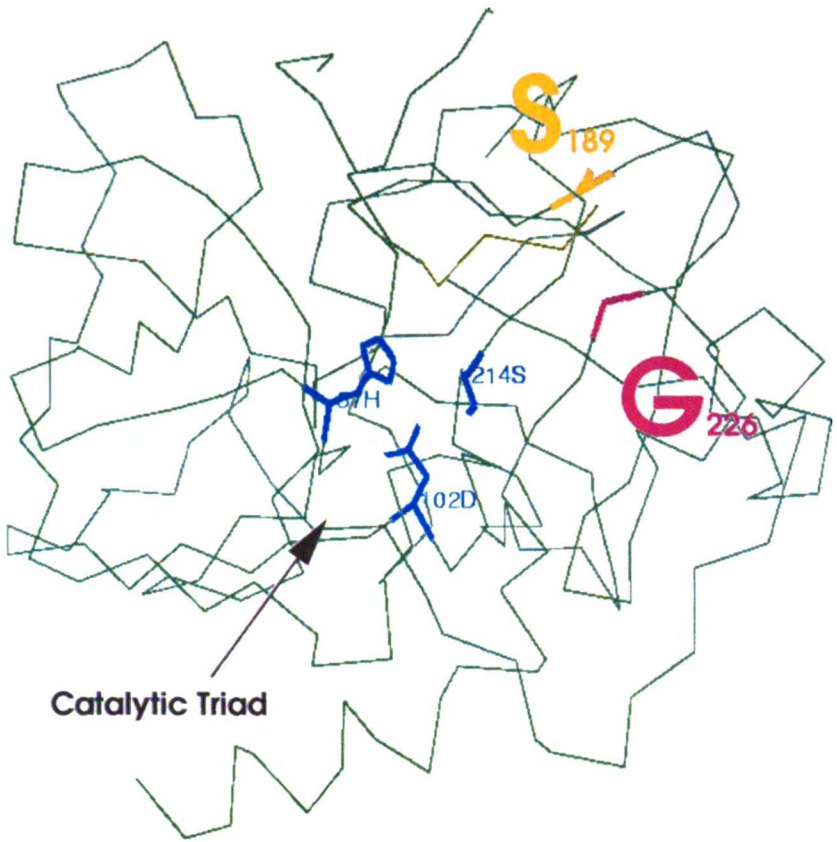


Figure Legend

a subclade of the phylogeny

invariance comparison between subclades

(X,Y) consensus sequence of a subclade

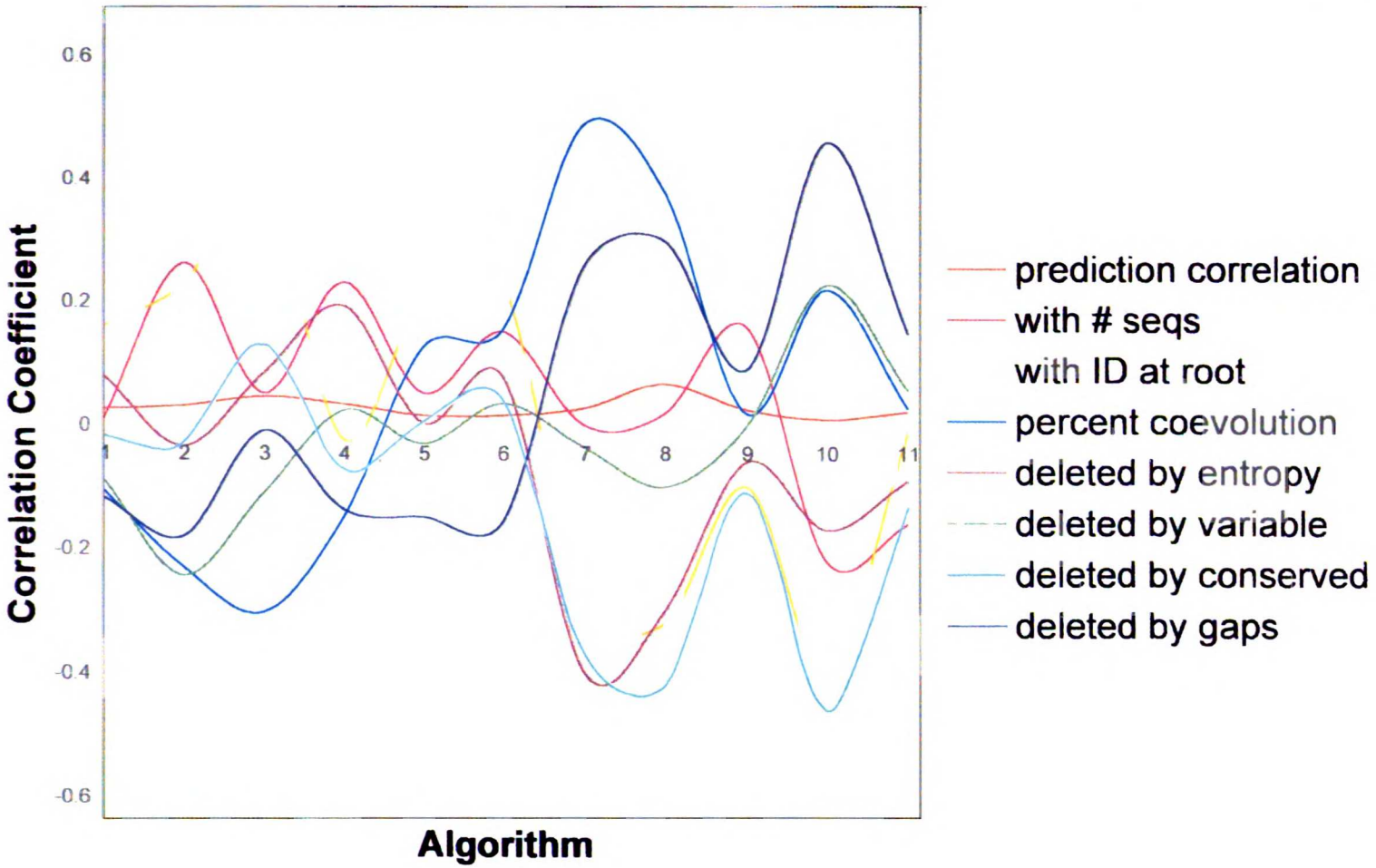
avoidance of problematic areas of the protein contact map, did not significantly improve the measures of success used to ascertain the contact information content in evolutionary data. Correlations to other parameters describing the protein family were tested, including the sequence identity at the hypothetical root of the phylogeny, the number of sequences in the family, and the overall percent of invariant positions. However, none of these feature filters exhibited clear trends applicable to evolutionary correlation based contact prediction (Figure 4).

Three measures of contact map prediction success were implemented (see Methods). The Matthews correlation coefficient (Matthews 1975) is useful when assessing the accuracy of prediction methods, since it takes into account both positive and negative, real and predicted information. The signal to noise ratio is a less sophisticated measure defined by the ratio of correct predicted contacts to the total number of predicted contacts. This ratio only addresses the absolute quality of the predicted contacts. Finally, the improvement over a random model, in this case actual protein structures, addresses the signal to noise problem in context of the contact information signal in real protein structures.

Figure 4

Results of tracking algorithm development by correlating prediction success with protein family features. 10 independent protein families were used for this analysis. Certain properties are anticorrelated, and this was found to be related to the information content in the evolutionary data. Specifically, diverse protein families are more amiable to subtraction of the background false positive correlation signal. Signal subtraction was performed with a variety of filters aimed to reduce uncorrelated positions and noncontacts. A number of these filters were incorporated into the final method (see Methods).

Correlating Contact Prediction Success with Protein Family Features



A summary of the numerical results for the early stages of the prediction algorithm formulation is shown in Table 1. Seven protein families were surveyed to observe the relations between sequence length and the number of actual contacts, with features of the evolutionary sequence correlations. The initial implementation of contact prediction mapped all pairwise evolutionary sequence correlations to a predicted protein contact map (Figure 5). Predicted contacts were evaluated by comparison to the actual protein contacts using three independent measures of success (see Methods). Importantly, in all cases the majority of real protein contacts were correlated in evolution (Figure 5). There were no persistent trends relating the number of actual contacts, with the number of evolutionary correlations, or the number of evolutionary residue contact correlations. However, certain protein families exhibited significant accuracies and improvements, namely the HIV proteases, the profilins and inosine monophosphate dehydrogenases (IMP) (Table 1). An avenue for improvement was seen in that every family contained more correlated residues than expected numbers of contacts in protein structures.

The primary problem in contact prediction from evolutionary correlations was the abundance of correlation

Table 1

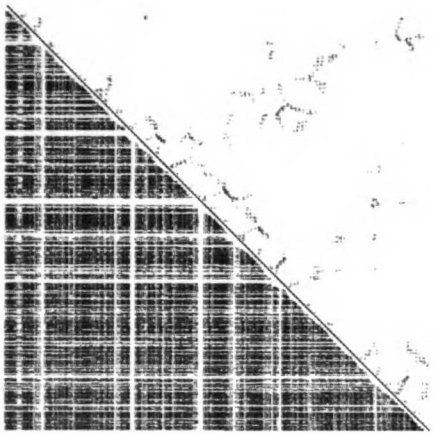
Results of contact prediction from raw evolutionary correlations. Nine protein family contact predictions were investigated with measures of accuracy (see Methods). Note the large number of predicted contacts compared to real contacts. At this stage of the prediction algorithm smaller proteins tended to have higher contact prediction accuracies.

Protein Family	Chain Length	Total Correlations	True Correlations	Real Contacts	Signal to Noise (%)	Accuracy (%)
lactalbumin	123	2734	64	200	2	32
lysozyme	129	3819	120	231	3	52
lactalbumin with lysozyme	129	3446	94	200	3	47
HIV protease	99	1227	281	680	23	41
creatine kinase	380	24803	1436	4235	6	34
aspartyl amino	822	37152	2109	5204	6	10
inosine monophosphate dehydrogenase	329	53787	6757	14424	12.6	57
profiliin	125	6295	544	990	9	55

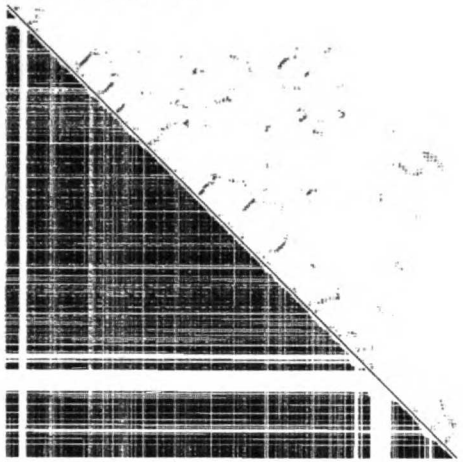
Figure 5

Contact predictions from raw evolutionary correlations. The upper diagonal of each contact map represents the real protein contacts. Correctly predicted contacts are in green and not predicted contacts are in magenta. The lower diagonal of the contact map represents the predicted contacts. The correlations scores are colored coded black to red, from lowest to highest evolutionary correlation score respectively. The abundance of evolutionary correlation data relative to real protein contacts can be seen. The measures of accuracy indicate that the predictions are with little correlation to the real contacts and close to random, and yet most real contacts are correlated.

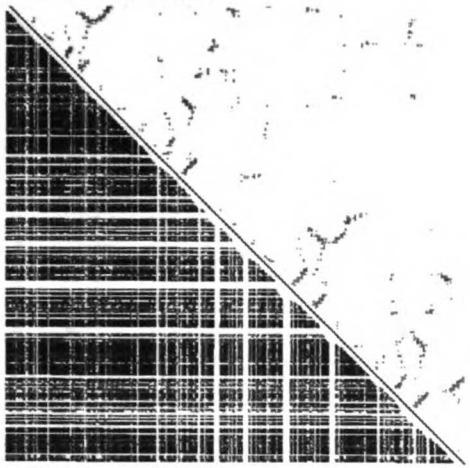
Asparyl Amino Transferase (1ART)
S/N: 1%
Improvement: .9
Correlation: -.02



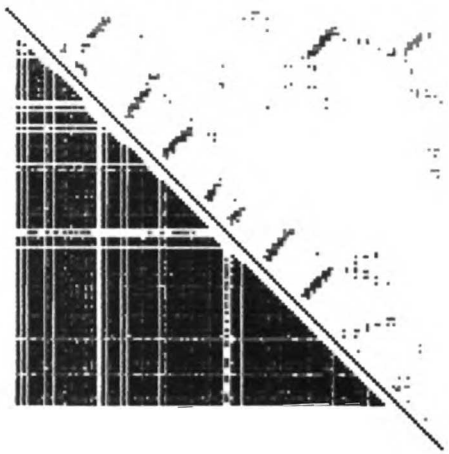
Catalase (2CAE)
S/N: 1%
Improvement: 1
Correlation: .01



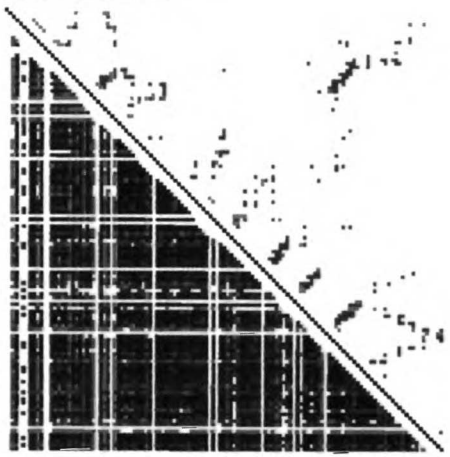
Biphenyl Extradiol Dioxygenase (1HAN)
S/N: 2%
Improvement: 1
Correlation: .004



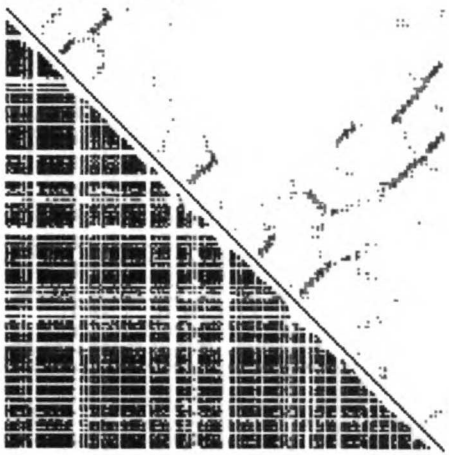
Odorant Binding Protein (1A3Y)
S/N: 4%
Improvement: 1
Correlation: .03



Profilin (2ACG)
S/N: 4 %
Improvement: 1
Correlation: .03



Xylanase (1BK1)
S/N: 4%
Improvement: 1
Correlation: .005



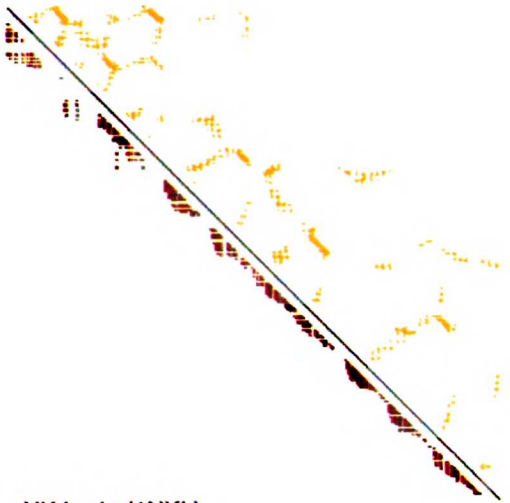
information relative to actual protein contacts. To address this problem and to increase consistency of prediction evaluation, the number of predicted contacts was set equal to the length of the protein sequence. A sorting scheme was implemented to collect the top correlations (see Methods). The results are significant in both signal to noise and improvement over random measurements (Figure 6). However, virtually all predicted contacts corresponded to local secondary structure contacts, with very few medium range interactions. Since the density of protein contact maps is greatest in the vicinity of the matrix diagonal, the chance of randomly predicting a correct local contact in this region is greater than in other areas of the map. This feature is an artifact of the sorting procedure as well as false positive correlations in the evolutionary data. Algorithmic sorting requires the assumption of a starting point. In order to reproduce natural qualities of contact maps, namely the decreasing density of contacts at larger sequence separations, sorting was performed beginning from the diagonal of the contact matrix, i.e. local contacts. This procedure biases the results to local and medium range contacts, which dominate all real contact maps.

To further address the problem of contact overprediction, a sequence separation based pairwise

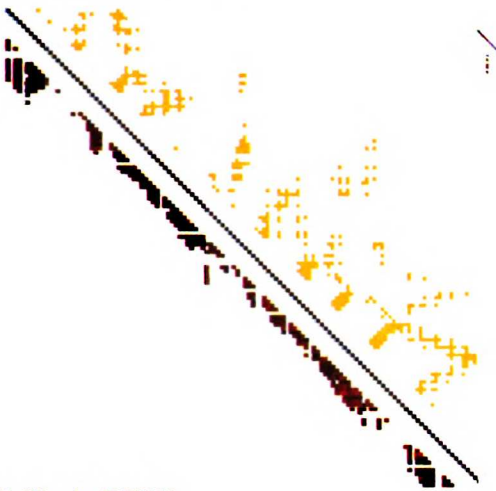
Figure 6

Contact predictions from limited and sorted evolutionary correlations. The upper diagonal of each contact map represents the real protein contacts. The lower diagonal of the contact map represents the predicted contacts. Contact predictions from evolutionary correlations were limited to a number of top correlations equal to the length of the protein. The correlations are sorted starting from the diagonal of the contact matrix (see Methods). There are also some improvements over the random model and in the signal measure.

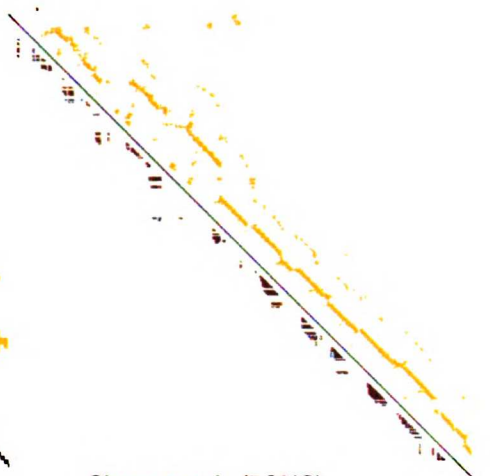
Hydroxysteroid dehydrogenase (1IOL)
S/N: 9.0 %
Improvement: 2.07



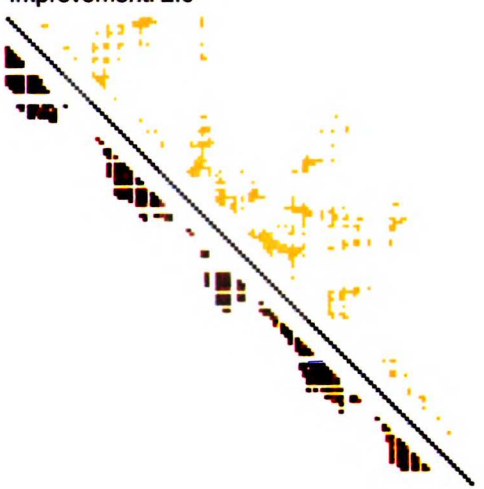
Profilin (2ACG)
S/N: 16.8 %
Improvement: 1.9



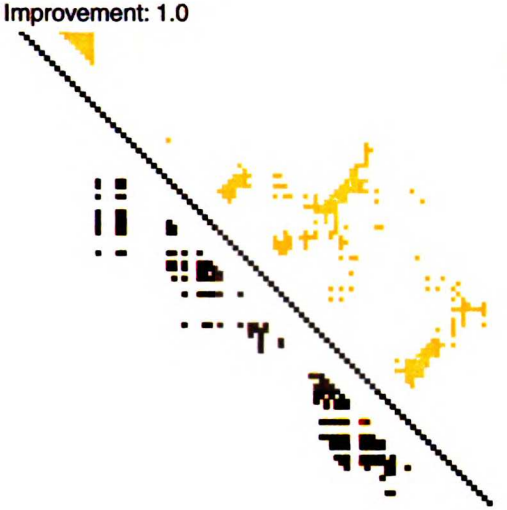
Polygalacturonase (1BHE)
S/N: 7.3 %
Improvement: 1.1



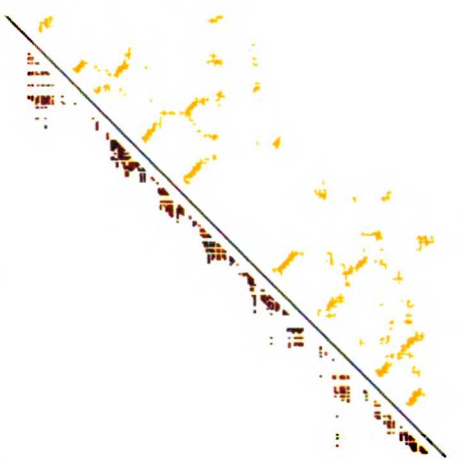
NK Lysin (1NKL)
S/N: 14.0 %
Improvement: 2.0



Kallikrein (2PKA)
S/N: 10.2 %
Improvement: 1.0



Chymotrypsin (7GHC)
S/N: 8.6 %
Improvement: 1.5



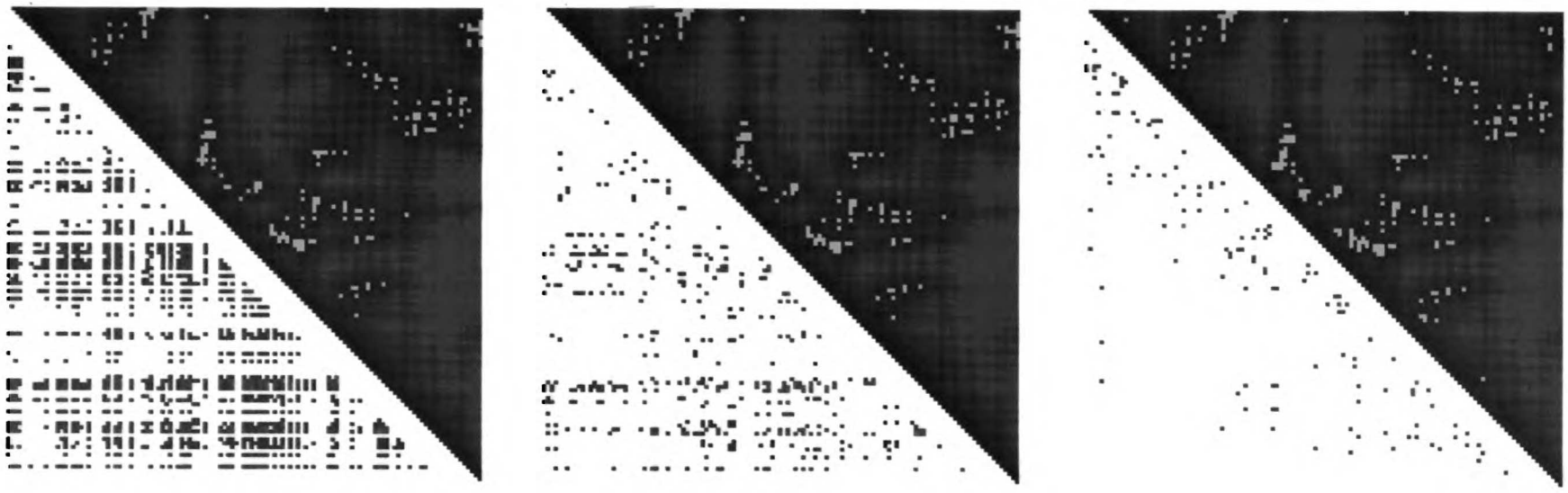
potential of mean force (Sippl 1990) was calculated based on a nonredundant subset of the PDB database (see Methods). This pairwise interaction potential was applied to the pairwise evolutionary correlations. With choice of an optimal minimal contact energy (see Methods), this procedure resulted in prediction of scattered contacts and less contact density (Figure 7). The scattering of predicted contact, especially at larger sequence separations, was in stark contrast to the multiple clusters of contacts visible in real contact maps (Figure 1).

In this method, problems in evolutionary data have direct effects on the nature and quality of protein contact predictions. Linearity of contacts, corresponding to rows and columns in the contact map, was the prominent feature of contact maps predicted from raw or interaction potential filtered evolutionary correlations. The lactalbumin family evolutionary correlations filtered with increasingly stringent minimum contact energy cutoffs illustrate this point (Figure 7). The evolutionary correlation information limited to a number of correlations equal to the sequence length could be reduced to a small set of residues correlated to many other residues. Many predicted residues in contact were correlated to indirect contacts potentially

Figure 7

Illustration of the conflicts in applying a pairwise interaction potential to contact map prediction. In this example, the lactalbumin family consensus contact map prediction was filtered successively with increasing pairwise interaction stringency. On the right the prediction approximates the raw evolutionary correlations. The number of predicted contacts in the middle prediction approaches the number of contacts in the real contact map, however no discernable protein contact map features are observed. The left contact map prediction with few scattered contacts represents a highly stringent application of the pairwise potential. In all cases the dominant prediction of contacts occurs in rows and columns of the contact map.

INCREASING NUMBER OF CONTACTS



INCREASING CUTOFF STRINGENCY

forming networks. Regardless, the majority of predicted contacts corresponded to false positive correlations.

The averaged results for 26 protein families are shown in Table 2. This prediction data was based on sorted evolutionary correlations filtered by the pairwise Sippl potential (Sippl 1990). An effect of a pairwise contact energy cutoff can be seen to have positive results on the correlated contact signal in the data. Invariably the linear cutoff approach presents the problem of predicting too few or too many contacts (Figure 7). Alternately, in the case of nonlinear cutoffs, decision making and approximations become problematic.

To overcome local contact prediction bias and to direct the sorting of correlations, occupancy constraints were applied to the evolutionary correlations filtered with the pairwise potential (see Methods). The occupancy constraints consisted of database derived frequencies of contacts in the local neighborhood of residues, using intervals of sequence separation as a discriminating parameter (Figure 8). Conditional occupancy restraints, relying on conditional probabilities of local residue contacts (see Methods) were generated and applied in a similar way (Figure 8). These distributions were used to

Table 2

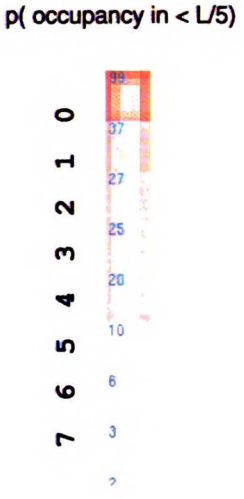
Contact prediction results for 26 nonredundant protein families. A small improvement in the correlation coefficient is seen when using a more stringent minimal contact energy cutoff. However, other measures of accuracy do not reflect this trend.

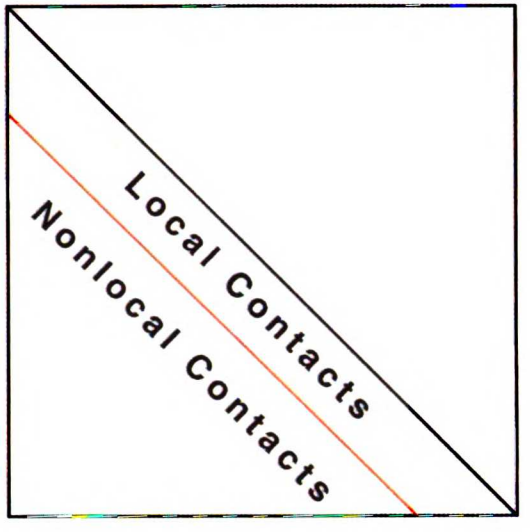
Prediction Method	Correlation Coefficient	Signal to Noise (%)	Improvement over Random	Correlated Pairs (%)
A : evolutionary correlations with occupancy constraints and limited sorting	0.01	2	1	80
A + pairwise potential with minimum energy = (L,T)	0.03	13	5	73
A + pairwise potential with minimum energy = average energy	0.05	9	4	75

Figure 8

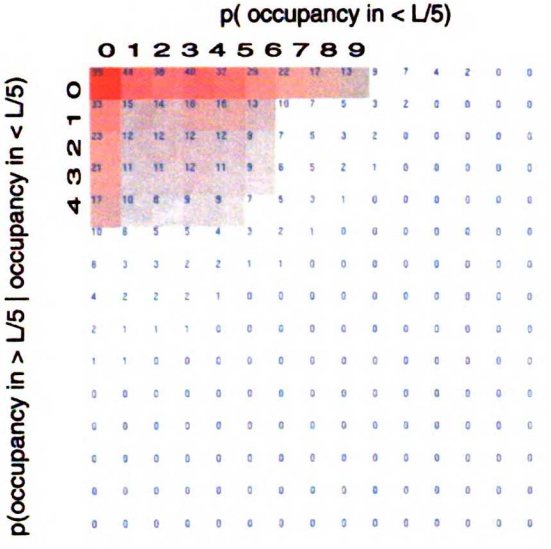
Occupancy constraints and prediction of protein contact maps. In the middle is a schematic of regions of the contact map corresponding to intervals of sequence separation. Local contacts were defined by a sequence separation interval from 4 to the minimum of $1/5^{th}$ of the chain length or 25 residues. Nonlocal contacts were defined by a sequence separation greater than the minimum of $1/5^{th}$ of the chain length or 25 residues. Unconditional local residue contact occupancy frequencies are shown on the left. Most residues in protein structures make no contacts. The conditional nonlocal contact occupancies for contact map prediction are shown the right. This knowledge-based data was used to conditionally assign nonlocal contacts from the prior (unconditional) assignment of local contacts (see Methods). Most residues have no contacts in the local sequence separation band. However, residues can participate in nonlocal contacts without forming local contacts.

Local Contact Probabilities





Nonlocal Contact Probabilities



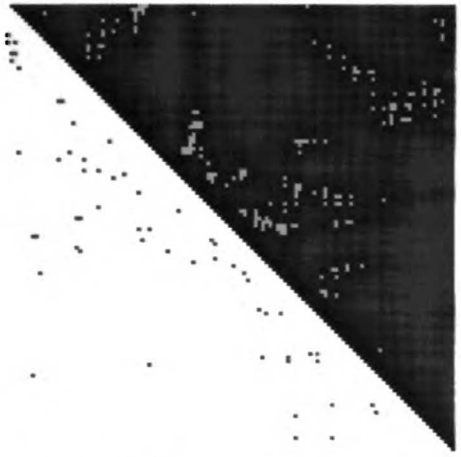
limit the number of contacts predicted per local neighborhood of a residue and to conditionally enforce the prediction of clusters of contacts based on prior predicted contacts. The results can be seen in Figure 9. Predicted contacts are dispersed throughout the contact map. The accuracy and signal to noise measure have decreased. Apart from loss of contact information, the degraded performance can be partially attributed to the fact that evolutionary residue contact correlations are not uniformly distributed across the parameter of sequence separation. In retrospect it appears necessary to assess the evolutionary signal for residue contacts in varying bands of sequence separation. Attempts were made to *a priori* quantify residue contact information content in the combination of phylogenetic and multiple sequence data. Difficulties appeared in forming parameters relating the phylogeny structure and sequence data to contact information across different protein families.

The final component of the contact prediction algorithm consisted of masking predicted contacts with patterns of contacts between secondary elements (see Methods). A consensus secondary structure for the protein family was predicted with the method of Chandonia and Karplus (Chandonia and Karplus 1999) (see Methods). A

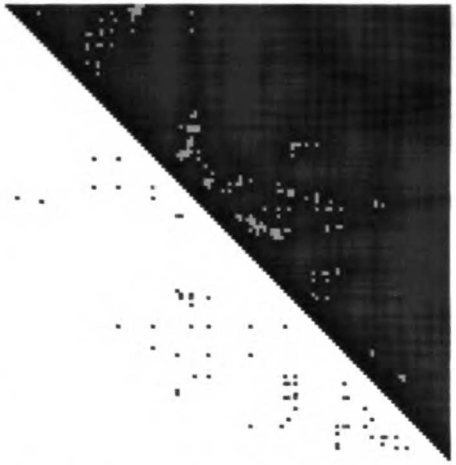
Figure 9

Application of a pairwise interaction potential and contact occupancy constraints to contact predictions from limited and sorted evolutionary correlations. The upper diagonal of each contact map represents the real protein contacts. The lower diagonal of the contact map represents the predicted contacts. See Methods for a description of the prediction procedure. Although the predicted contacts are scattered, they cover all sequence separations in the contact map. The measures of prediction success indicate a significant improvement due to the pairwise potential and contact occupancy constraints.

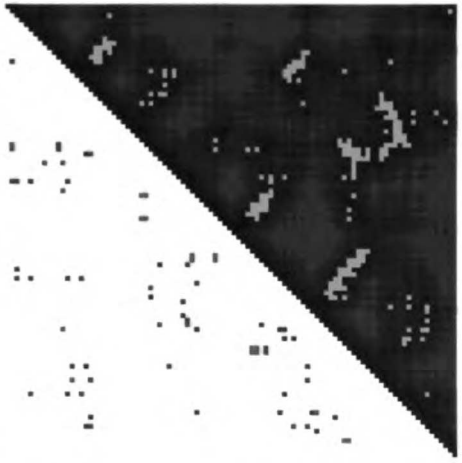
Lactalbumin (1HML)
S/N: 9.0%
Improvement: 2.6



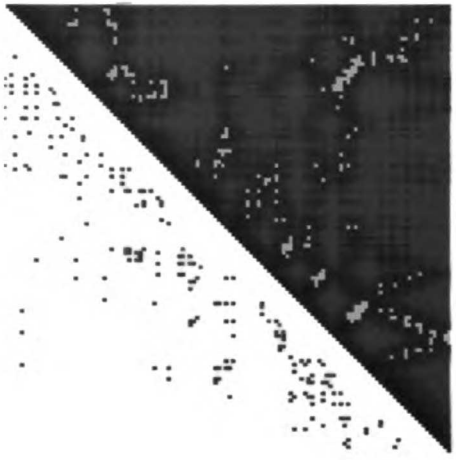
NK Lysin (1NKL)
S/N: 8.0%
Improvement: 3.0



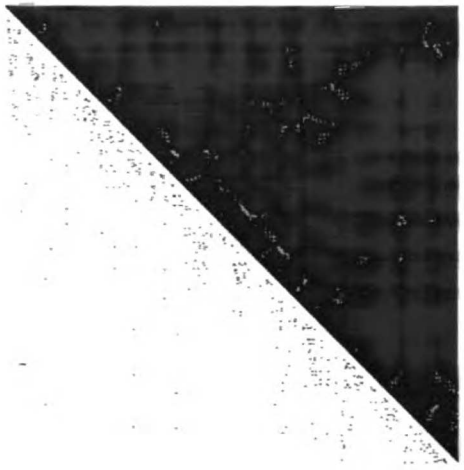
HIV protease (1A94)
S/N: 12.0%
Improvement: 2.7



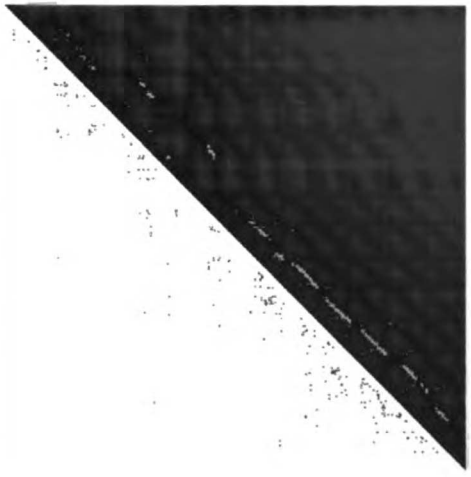
Profilin (2ACG)
S/N: 7.5%
Improvement: 2.6



Amino Aspartyl Transferase (1ART)
S/N: 2.5%
Improvement: 2.4



Polygalacturonase (1BHE)
S/N: 13.0%
Improvement: 5.3



mapping of the secondary structure prediction was used to identify the potential interaction spaces of secondary structure elements. Iterative masking of unconditional and conditional secondary structure interaction probabilities was driven by data from previous layers of the algorithm (see Methods). This procedure resulted in predicted contact maps with features reminiscent of real protein structure (Figure 10). In particular, clusters of contacting residues attributable to helix-helix, helix-beta strand, and beta strand-beta strand interactions were observed. However, certain families fared worse with this method than with the combination of the pairwise potential and occupancy constraints. Further work was needed to complete the iterative procedures in the algorithm and to cohesively analyze the performance and limitations on a larger test set of protein families. In addition, due to a large number of variable parameters in the method, further optimization of parameters in light of the measures of contact prediction success was necessary.

The contact maps predicted by the final algorithm exhibited features reminiscent of real protein contact maps. Prediction of such contact map features based on sequence correlations has not been reported thus far. In fact, authors of related prediction methods commented on

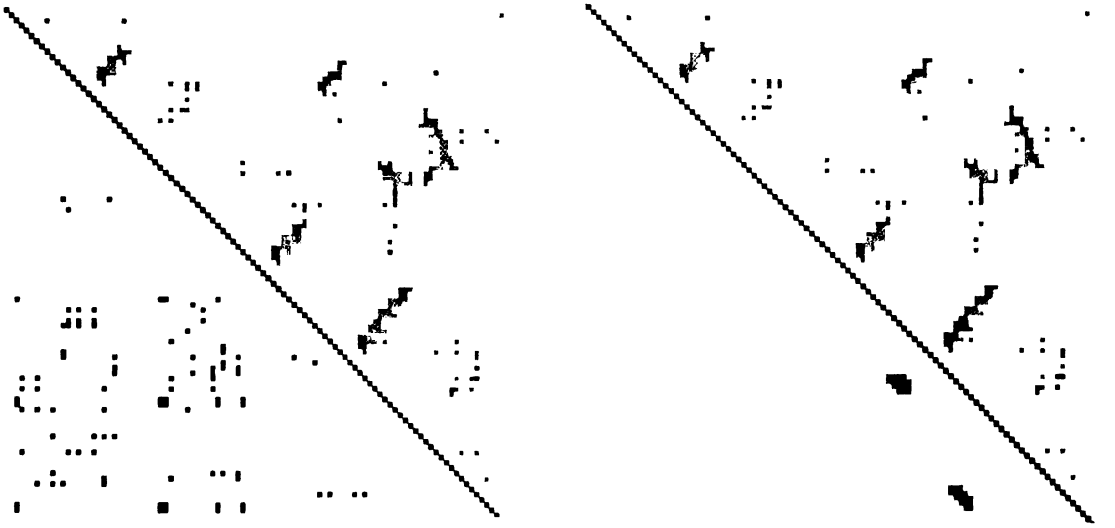
Figure 10

Contact prediction with the final algorithm. The procedure in these predictions included limited and sorted evolutionary correlations, contact occupancy constraints, a pairwise interaction potential, predicted secondary structure and conditional inter-secondary structure interaction probabilities. **A** demonstrates the noticeable improvements achieved with addition of predicted secondary structure and conditional inter-secondary structure interaction probabilities to the prediction method. **B** shows the best prediction, the consensus contact map of the profilin family. Although the correlation coefficient was still relatively small, the signal to noise and improvement over random models indicate a potentially useful structure prediction.

A

Raw evolutionary correlations
HIV protease (1A94)
S/N: 11.4 %
Improvement: 2.4
Correlation: 0.05

Final algorithm result
HIV protease (1A94)
S/N: 34.4 %
Improvement: 7.3
Correlation: 0.12



B

Profilin (2ACG)
S/N: 40 %
Improvement: 10
Correlation: 0.2

the scattered character of contact predictions from evolutionary correlations (Casari, Sander et al. 1995; Olmea and Valencia 1997; Olmea, Rost et al. 1999). It has to be noted, that under the constraint of a limited number of residue-residue contacts, prediction of scattered contacts results in a higher chance of random correct prediction. In this light, it is a more significant achievement that the present method succeeded in reconstructing features of protein contact maps from evolutionary sequence correlations and knowledge based structural constraints.

Discussion

The main advantage of phylogenetic information is that it can represent biologically meaningful sets of residues correlated between sequences, or groups of sequences. The precise goal was to use the structure of protein family phylogeny to identify contacting residues. The underlying assumption was that the evolutionary processes of selection and mutation are evident in the patterns of phylogenetic subclade divergence (Lichtarge, Bourne et al. 1996). Using an abstract definition of evolutionary time related to measures of sequence distance (Dayhoff 1972), the phylogenetic structure presents a historical context for modern sequences. These concepts introduce a mathematical correlation based on phylogenetic structure and multiple sequence data.

Gathering both survey and test input data for the method was time consuming and nontrivial. The number of protein structures of sufficient quality and coverage of the protein sequence was the first limiting step. Abundant sequence data was the second rate limiting step. For any protein family satisfying the minimal information requirements, a correlated contact signal larger than

randomly expected was usually present. This result was promising for extracting sequence correlations related to protein structure, especially in combination with results supporting the positive effect of increased evolutionary diversity in a family on the pairwise correlation signal (Pollock and Taylor 1997; Pollock, Taylor et al. 1999).

The presented method relies on iterative subtraction and addition of signals coming from evolutionary pressures on amino acid sequence, residue-residue interactions, secondary structure, and from tertiary interactions of secondary structure elements. This data is fundamentally related to the subclade structure of the phylogeny, since the initial prediction of contacts relies on a concept of evolutionary correlation. The knowledge-based potential of Sippl (Sippl 1990) in a somewhat modified version (see Methods), was applied prior to each step of the iterative procedure to ensure that only favorable interactions were considered for each iteration.

Of all components of the algorithm, the introduction of conditional secondary structure interaction probabilities provided the single largest increase in prediction success. Secondary structure can be predicted with high degrees of accuracy from multiple sequence data (86 %, (Chandonia and Karplus 1999)) and a portion of the

contact prediction improvement resulted from biasing contact prediction with predicted secondary structure. However, in cases of "nearly" perfect prediction by current standards (Figure 10B), the improvement came from correctly assigning a cluster of contacting residues by the conditional secondary structure contact procedure. The Bayesian statistics used to iteratively assign secondary structure contacts contributed significantly to the overall contact prediction procedure. The use of secondary structure masks magnified the residue contact signal in the evolutionary correlations, achieving higher levels of prediction accuracy than evolutionary data alone or in combination with a variety of filters. Additionally, secondary structure probabilities resulted in the tolerance of contact prediction to shifts in register, twists of secondary structures and small irregularities such as kinks in beta sheets and helices within a protein family. Many such protein structure irregularities lead to problems in sequence alignment and hence loss of evolutionary information. It appears that using secondary structure contact probabilities to assign contacts based on evolutionary data input, allows regaining some of the lost information due to alignment and phylogenetic reconstruction errors and structural divergence.

Independently, masking contact maps using predicted secondary structure provided a useful analysis of the nature and contributions of local and non-local secondary structure contact correlations.

Correlations derived from evolutionary sequence data contain significant background noise. Sequence changes that appear correlated but are distant in the protein structure, can occur in a variety of situations. These include random amino acid mutations that cannot be distinguished from compensatory mutations of residues proximal in space. Such random changes have none or little selective consequence, and yet they are inherited in similar patterns as correlated contacts. Irrespective of the random background in evolution, there are specific patterns of sequence variation indicative of evolutionary selection.

A number of correlations can be biologically meaningful, even if the involved residues are distant in space. Members of a protein family that have multiple functional sites fall into this category. Pollock et al 1999 have suggested that a component of the 'random' background is related to maintenance of protein surface chemical properties, ultimately related to phenomenon such as compatibility of secondary structures and oligomerization avoidance. It seems that such positions

should be distinguished by recognizable patterns in protein families, relative to random drift. Many limits in interpretation of the evolutionary nature of sequence changes can be traced to evolutionary events poorly modeled with accepted phylogenetic methods. Rapid speciation, variable rates of substitution, variable amino acid preferences, and possible functional and structural consequences of different amino acids in different species and environments, all present problems for reconstructing accurate phylogenies. In particular, nucleotide and amino acid bias contribute recognized errors to phylogenetic reconstruction. Especially significant is the AT/GC bias at the nucleotide level (Foster and Hickey 1999). The case of unique homologs to different regions of a sequence presents an especially difficult case for current methods. Unfortunately, alternative interpretations of sequence evolution are not easily amenable to algorithmic incorporation. Computational challenges are present both in numerical interpretations such as singular value decomposition, which is difficult to relate to biological context, and more biological-like models whose complexity introduces implementation and performance issues (Devauchelle, Grossmann et al. 2001; Kitazoe, Kurihara et al. 2001). Nevertheless, there are some interesting methods

to assess phylogenetic content of multiple sequence data (Wagele and Rodding 1998; Strimmer and Moulton 2000). Incorporation of these techniques would enable more precise dissection of evolutionary correlations.

Positions that are conserved in a protein family have a zero information content with respect to pairwise evolutionary correlations. Yet it is a given that positions invariant in the hypothetical ancestor can be mutated without effects on structure and function. In the case of known protein structures, it is routine to assign contexts to absolutely conserved positions through protein family sequence conservation patterns. Knowledge of the determinants of a protein's function can be an asset in interpreting correlations. Biasing contact predictions by the presence of conserved residues within a short sequence separation would be expected to increase structure prediction accuracy measures. Incorporation of the significance of conserved residues into a contact map prediction method could be also achieved by using available functional information to assign chiralities or even approximate orientations for conserved residues. Preliminary attempts to incorporate functional information have aided *ab initio* structure prediction (Benner, Gerloff

et al. 1995; Gerloff, Joachimiak et al. 1998; Gerloff, Cannarozzi et al. 1999).

To find correlated positions, the residue changes that induce residue changes due to evolutionary selection criteria, it is possible to use data other than amino acid sequence. The use of a ratio of synonymous to nonsynonymous gene substitutions, as derived from the codon translation table, provides another layer of evolutionary data for protein sequence and structure analysis (Liberles, Schreiber et al. 2001). The ratio of synonymous to nonsynonymous mutations has been applied to identify potential cases of functional change (Li, Wu et al. 1985; Messier and Stewart 1997; Yang, Nielsen et al. 2000). However, it appears that a Hidden Markov Model for variable substitution rates in subclades of a phylogeny (Gu 1999; Gu 2001; Gu 2001), more accurately describes the sequence variability indicative of evolutionary selection for function (Gaucher, Miyamoto et al. 2001). It remains unclear which method is most suitable for specific gene families (Gu 1999). Introduction of nucleotide sequence input data for contact prediction, introduces a new set of evolutionary pressures potentially unrelated to protein sequence, structure and function. Thermophilic requirements for DNA stability which result in a guanine and cytosine

nucleotide bias, and the unknown reasons for adenine and thymine nucleotide richness of the malaria genome (Weber 1987) also effect protein sequence evolution. More difficulties reside in the nonuniversality of the genetic code and codon frequencies, which influence amino acid composition (Foster and Hickey 1999). However, the synonymous to nonsynonymous substitution ratio is an interesting measure, since over a family of protein sequences, the biases of the genetic code succumb to the statistics of large numbers.

A function extracting information from the connected layers of information present in DNA and protein sequence data could provide information about random drift positions. One approach is to take the sum of the number of unique codons for every amino acid at an alignment position, and to divide it by the number of invariant subclades at that position. This measure can be applied at different evolutionary distances within the phylogeny, based on partitions of the subclades (see Chapter III). Further development was necessary to fully integrate nucleotide related segments of the prediction algorithm.

A number of additional modifications to the final algorithm are possible and potentially useful. Contact types should be discriminated at the atomic level and more

closely related to the chemistry of protein structure and function. The amino acids arranged in a similarity tree (Smith and Smith 1992), are enticing for constructing reduced alphabets and interpreting evolutionary correlations through a physical-chemical perspective. Another idea is to overlay phylogenetic trees representing different properties: a variety of substitution matrices, reduced amino acid alphabets, predicted secondary structure, ratio of synonymous to nonsynonymous nucleotide substitutions. These approaches could increase the information content of pairwise evolutionary correlations. In general, combinations of orthogonal data lead to multiplication of associated errors, compared to the addition of errors when combining dependent information. In this case algebra minimizes the effects of errors, and theoretically high reliability can be achieve from data with weak signals.

In a novel survey of sequence to residue contact correlations, Selbig and Argos argue that introduction of explicit noncontact information could improve the accuracy of contact prediction (Selbig and Argos 1998). The authors constructed a model of correlations using a number of contact environments. Based on pairs of sequence triplets they showed efficient discrimination of contacts and

noncontacts. However, the authors concede that the decision tree approach to clustering contact types was method and family dependent. Although evolutionary correlations contain a false positive signal, the data contains significant residue contact information. Combinations of this data with a pairwise potential, a strong discriminator of contacts and noncontacts, and especially conditional limitations on the number and density of contacts, results in a reasonable incorporation of noncontact information into this contact prediction method.

A significant number of accurate predictions were made based on the raw evolutionary correlations and were improved with a number of procedures, namely applying criteria of interaction energy and predicting supra-secondary structure interactions. This suggests adequate detail in the residue-residue contact model in this method. Using distances between Cα atoms of the protein backbone introduces the problem of localizing the interaction in space. In this regard, the Cβ atom distances or even better centroids of sidechains represent a more detailed interpretation of protein sidechain-sidechain interactions. However, the contact model should also include the chemical interaction units of the protein termini and the protein backbone. These changes to the contact model can be

profound for the calculation of the amino acid pair potential.

The pairwise interaction potential could also be improved. The definition of contacts has a large influence on the calculated energies of residue contacts. However, a number of additional conditions could be differentiated in the potential calculation. The occupancy of the residues in the interaction pair would encode the packing and potential contact network information. The chain orientation of the interaction pair could also represent important differences in pair interaction frequencies, related to chirality and global architecture. Finally, the chirality of the backbone may play an important role in defining optimal contact types. Protein structure chirality can be independently defined by calculating the per residue chirality relative to the most N terminal residue. Statistical potentials have been criticized (Thomas and Dill 1996) as to their information content, and because they do not explicitly include the multiple dependencies characteristic of protein structure. In spite of these drawbacks, pairwise statistical potentials are popular entrees in the cookbook of computation biology procedures. The proposed changes to the calculation of the potential mean force (as first proposed by M. Sippl), address issues of information

content and suggest incorporation of parameters describing the coordination and multiple contact dependency of residues in protein structures.

Among input data requirements of the prediction method are sequence alignments maximizing global and local sequence homology. This maximization relies on the meaningful placement of insertions and deletions in sequence alignments. In survey experiments or algorithm development, the ideal sequence data corresponds to structural alignments. Unfortunately at the time of this work there were few protein families diversely represented by structural data. Since the resulting correlation data is related to the number of sequences in a nonlinear fashion, large protein families are preferable. This is to assure that the phylogenetic dendogram representing the given protein family corresponds to a minimal evolutionary diversity, usually a span of 30% in sequence identity (e.g. 70% identical residues in the hypothetical ancestor to 100% in individual sequences). A combined measure of structure mutability and sequence diversity is the frequency of conserved residues (or conserved contacts). Such measures would help characterize differences among protein families and the resulting correspondence of prediction heuristics. To ensure consistency between the phylogeny and multiple

sequence data, families of sequences with homology to multiple closest homologs in different regions of the sequence had to be excluded from the analysis. Unfortunately, the standard phylogenetic methods cannot represent such relationships meaningfully for structure prediction. In the test case of a known structure the contacts for a given residue were mapped to the MSA data to assess and analyze the prediction (see Methods). Future implementations were to iteratively apply structural data to improve the identification of correlated positions.

The number of contacts in a protein increases linearly with the sequence length, but the number of noncontacts increases as the square of the sequence length (Vendruscolo, Kussell et al. 1997). This result was independent of the residue contact definition. Statistically this implies that contact prediction will be increasingly difficult for larger proteins since the random chance of predicting a noncontact increases faster than the random chance of predicting a contact. In addition, Fariselli et al (Fariselli, Olmea et al. 2001) report that 81% of contacts occur in the sequence separation interval of 7 to 100 residues, and at larger sequence separations the contacts became more scattered and less clustered. These recently discovered protein contact heuristics

provide crucial information for structure prediction methods. Novel protein structure heuristics are required to ascertain the minimum and maximum expected number of inter-secondary structure contacts.

The number of parameters necessary to describe amino acids and their interactions in proteins is quite large. A common approach is the parametrization of Cα distances with a six term functions (Reese, Lund et al. 1996). Complex parameterizations, including neural networks, appear as a fundamental problem in protein structure prediction. The task of extracting heuristics for prediction improvement is further complicated by limited experimental data on the nuances of sequence to structure relationships.

Protein structure prediction based on sequence data is highly sensitive to errors and ambiguities in alignments. Sequence databases contain multiple types of errors and anomalies, and this is especially true of the sequence termini (Lamperti, Kittelberger et al. 1992). Prediction of residue-residue contacts in the N and C-termini of the protein are therefore especially unreliable. This effect has serious implications for short sequences and proteins with prominent secondary structure in the terminal regions.

The plasticity of protein structure has been documented in a number of settings. Notably, a shift in

residue register either within a secondary structure element or the register of interaction between secondary structures (Vetter, Baase et al. 1996), represents a significant mode of structural plasticity. It has to be recognized, that shifts in register often correspond to sequence insertion and deletion events. Such events are common within a protein family, especially in loops and regions not directly constrained by functional or structural requirements (Chapter II). Shift of register directly contributes to errors in contact prediction, since even a correct sequence alignment corresponds to an error in register. A related phenomenon is that gapped alignments can result in the scattering of predicted contacts in areas of an expected contact cluster. This problem is often related to difficulties in constructing a correct sequence alignment. A useful measure to assess the long range predicted contact information relies on contact specificity - a ratio of predicted or implied contacts relative to the actual protein structure (Marchler-Bauer and Bryant 1997). Such measures could enable scrutiny of errors in predicted contacts resulting from sequence alignment and structural irregularities.

An error function for protein structure prediction is an intensely complex matter. In fact, this problem is one

of the major reasons why four years later there are still no major advances in the protein structure prediction field, aside from combinatorial knowledge-based methods (Simons, Kooperberg et al. 1997). Ultimately an error function for structure prediction based on evolutionary correlations would include the probability of alignment errors (structural nonequivalence of positions), errors in the phylogeny from variable sequence substitution rates, and the associated errors for each additional technique used as a prediction layer or filter. Gaucher et al reported use of variable substitution rate methodologies to analyze structure and function in an example protein family (Gaucher, Miyamoto et al. 2001). One solution to the problem of assessing sequence uniqueness and variability distribution across the phylogeny is to define a phylogenetic information entropy. Such a function would encode distances between nodes in the tree, both in the distance from a hypothetical ancestor and the number of sequences in between a pair of sequences, as well as a ratio of the number of sequences to the number of branch points in the phylogeny. This description of the phylogenetic information content can be related to percent sequence identity or other sequence distance measures.

Identification of correlated residues provided combinatorial sets of possible contacts. Specific predicted interactions could be verified by threading, thus making use of independent prediction methods. Experimental verification in the form of site directed mutagenesis is another route to independently confirm predicted interactions or structures. An experimentally determined atomic structure is the best standard for assessing protein structure prediction, for example using a self-threading test (Orengo, Bray et al. 1999). A number of structure prediction methods attempt solutions to specific instances of protein structure or sequence to structure correlations. The prediction of disulfide bond connectivity (Fariselli and Casadio 2001) and prediction of residue coordination number (Fariselli and Casadio 2000; Fariselli and Casadio 2001) are examples of methods suitable for integration with this structure prediction method. New approaches to low-resolution structure determination provide orthogonal distance constraint information. Young et al (Young, Tang et al. 2000) report efficient fold identification utilizing intramolecular crosslinking and mass spectrometry. Combination of distance constraints from independent approaches could enhance the specificity and resolution of predicted structures.

Automated verification of the approach could consist of solving structures based on predicted distance constraints using distance geometry. The DGEOM software (Blaney, Crippen et al. 1984-1994) is a standard tool in NMR structure determination (Liu, Zhao et al. 1992). The existence of such methods ensures that with the required number of distance constraints, the true structure can be determined. Ensembles of structures produced by DGEOM can be analyzed to elucidate over and determined regions, adding to the iterative process.

The persistent problem in computational protein structure prediction is the inability to assess quality of *ab initio* predictions in absence of experimentally determined structures. Moreover, any computational efforts in this area are dotted with assumptions and approximations. Gobel et al (Gobel, Sander et al. 1994) assess prediction accuracy by randomizing protein sequences, and hence their contact maps. By randomizing the distinct features of protein contact maps attributable to secondary structure elements and their interactions, literal randomization produces contacts that are not indicative of realistic protein tertiary structure (Figure 11). Thomas et al (Thomas, Casari et al. 1996) use weights based on the structure being predicted as well as various

"estimates of unconditional probabilities" to normalize the predicted contact data. Selbig and Argos (Selbig and Argos 1998) analyzed sequence to structure correlations using definitions of types of contact environments and a three residue window, applying these results to contact prediction. The variation of models and methods implementing sequence to structure correlations for contact prediction, and the complexities of structure prediction and sequence and phylogenetic data analysis, make comparison of predictions and methods a daunting task. A corollary of the variation in data representation and prediction algorithms is that it is nontrivial to translate any resulting heuristics to other models and implementations.

In addition to numerical evaluation of prediction contacts, a visual measure based on the graphics of the Venn diagram (John Venn, British logician, 1834-1923) would be helpful. The overlaps between different sets of residues represent the common set of residues, aiding in discrimination of methods or parameters that lead to orthogonal predictive information (Figure 12).

Figure 11

Example of sequence randomization used in some contact prediction methods. The upper diagonal corresponds to the real contact map of profilin, while the lower diagonal represents the randomized contacts. The sequence of the reference protein structure, profilin (PDB code: 2ACG) was randomized using a random number generator for 1000 iterations. Unfortunately, the numerical solution to graphically represent overlaps of multiple geometrical objects is nontrivial and was not automated.

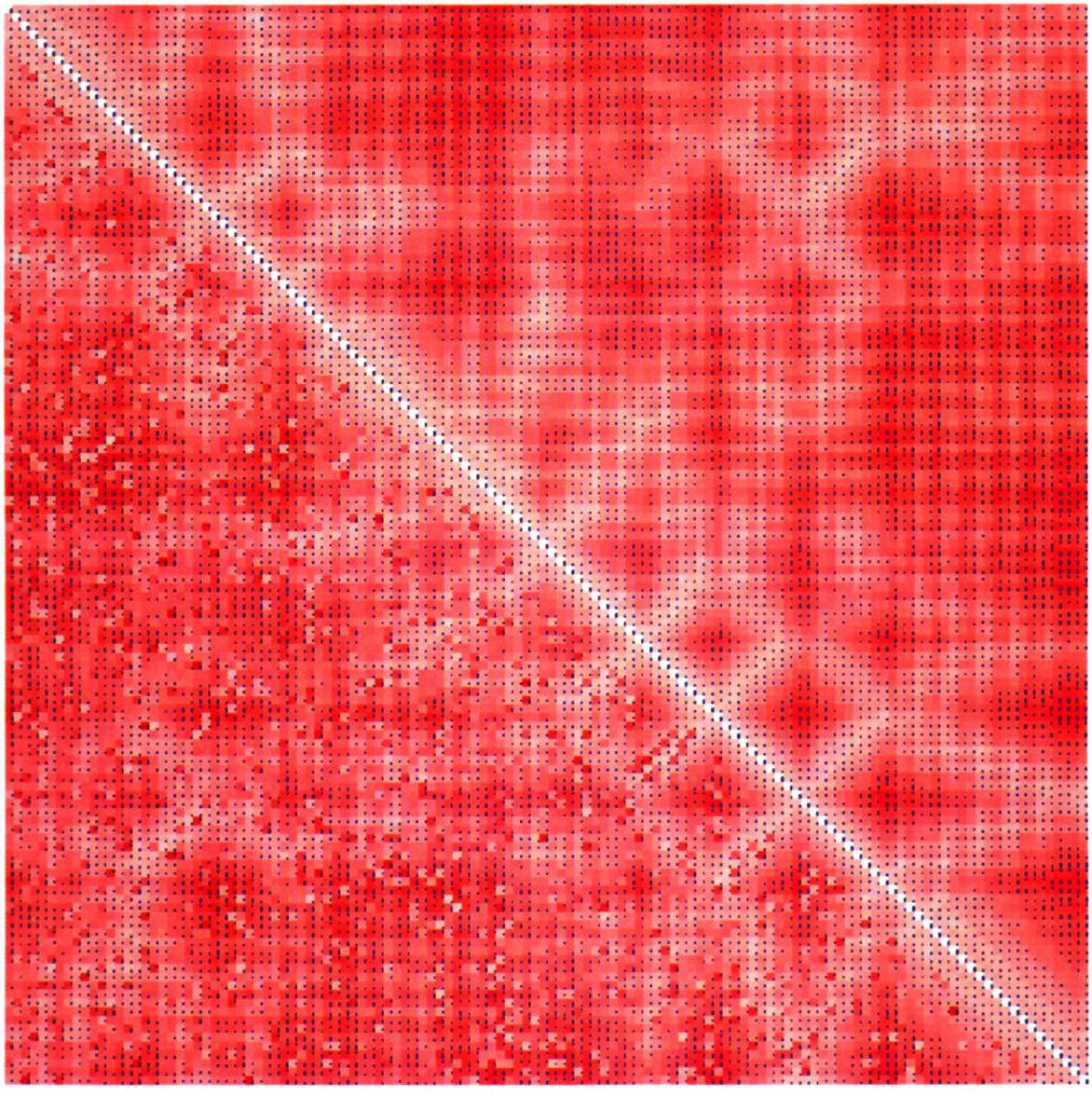
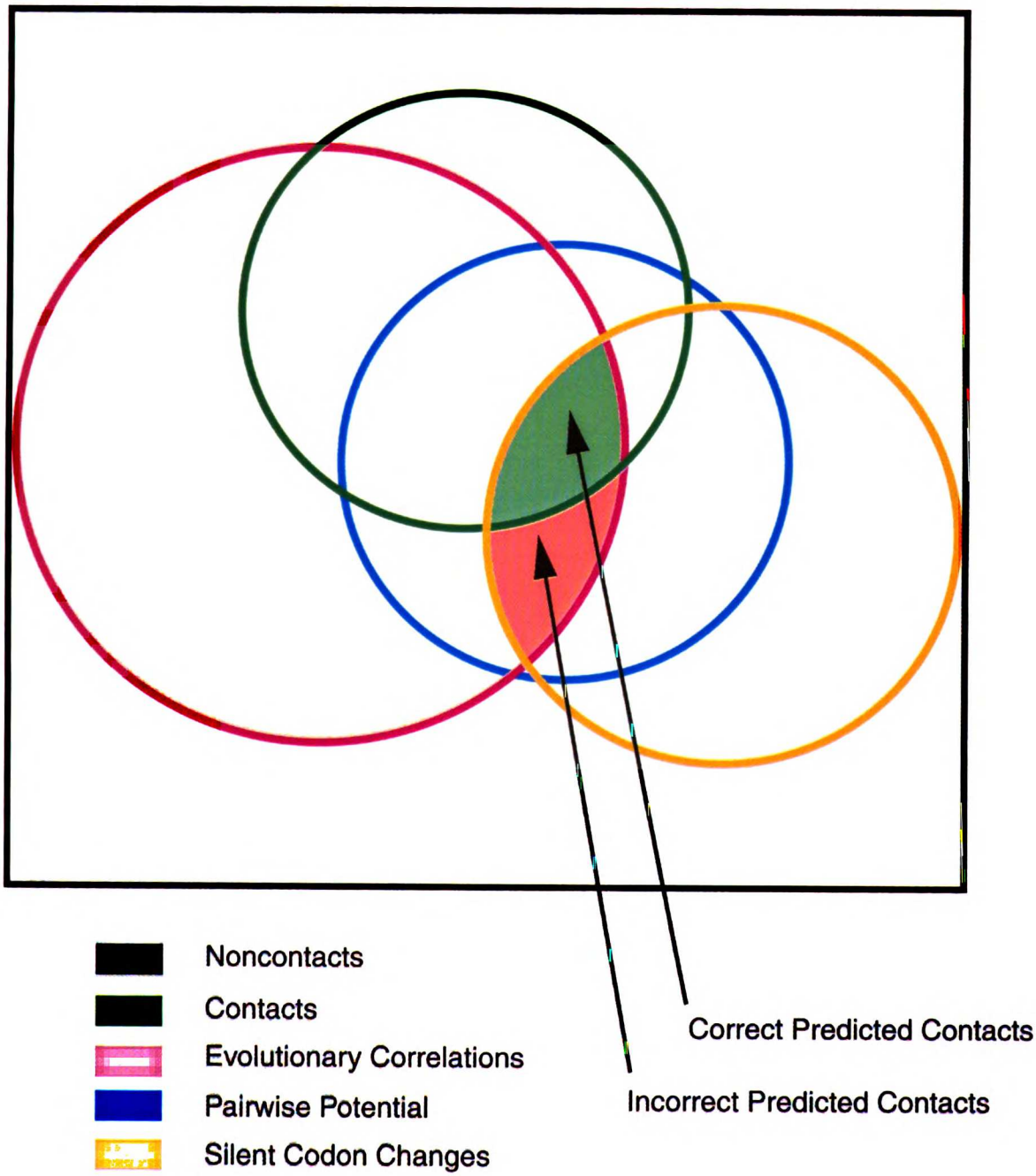


Figure 12

Example of a Venn diagram interpretation of contact prediction results. Colored circles correspond to different components of the prediction method. The shaded areas correspond to intersections of contacts predicted with these components. This representation is helpful in assessing the orthogonal contributions of data and methods to structure prediction.


Noncontacts
Contacts
Evolutionary Correlations
Pairwise Potential
Silent Codon Changes
Correct Predicted Contacts
Incorrect Predicted Contacts

The phylogenetic contact prediction approach represents a biologically meaningful normalization of the sequence correlation data, using the structure of subclades instead of numerical procedures. A number of methods resort to known distributions of contacts and noncontacts (Olmea and Valencia 1997; Pollock and Taylor 1997; Pollock, Taylor et al. 1999), however in this case strong limitations derived from the contact and correlation models as well as sufficient sampling of the known distributions, present problems for statistical discrimination in contact prediction. Neural networks method reported by Fariselli et al (Fariselli, Olmea et al. 2001), in addition to the compounded issue of parametrization, do not provide transparency for iterative analysis of the input data and results. Iterative improvement appears crucial in developing and tuning complex algorithmic procedures. Iteration proves to be more than a heuristic and parameter 'debugging' utility. Analysis of structure predictions in light of real protein structures enlightens the underlying patterns in sequence to structure correlations and allows formulation of novel structure prediction heuristics.

Conclusions

Biological mutability is fundamentally related to the evolution of structure and function. The genetic code and cellular mechanisms of replication, transcription and translation are optimized under evolutionary constraints to ensure stability of function, and hence structure. These constraints limit divergence within a phylogeny, while effects of the biological optimization are evident in the patterns of sequence changes within a family. The initial surveys in this work demonstrated that evolutionary correlations contain a significant signal relating to residue-residue contact information. Inseparable from this correlated contact signal was a false-positive background of complex evolutionary events, including neutral sequence drift. Protein contact prediction attempts the balance of realistic protein structure, which is intrinsically complex, with equally complex input data that benefits from large numbers of parameters and sample sizes due to background noise. One of the most serious difficulties in these prediction attempts are limitations on assessing the prediction accuracy, especially in absence of experimentally determined protein structure.

Complexity of alignments and significant variability between families with respect to phylogenetic structures were considerable limitations in surveying protein families for correlated contacts. These limitations reappeared in the structure prediction algorithm and prediction improvements approaches. Just four years later, the amount of sequence, function and structure data has increased dramatically. In addition, there have been significant improvements in bioinformatics methods, notably database sequence searches, fold recognition and structure comparisons. Arguably even greater improvements have occurred in the realms of protein characterization and structure determination. Importantly, many of these advances are being coupled to the post-genomic environment in modern biology. Currently all sequences with structure homologs have been modeled and annotated (Pieper, Eswar et al. 2002), a detailed database of protein and gene evolutionary families has been developed (Liberles, Schreiber et al. 2001) and the first threading of an entire genome is reported (JMC unpublished - *Drosophila melanogaster*). Last but not least, structural genomics initiatives based on novel fold approaches are increasing the pace of structure deposition (Stevens, Yokoyama et al. 2001). Although the protein folding protein remains

unsolved and protein structure prediction unsatisfactory, a number of different avenues are leading to protein structure or structure-related information.

Sequence to structure relationships are still tedious to dissect, involving directed sequence evolution (Raillard, Krebber et al. 2001; Sieber, Martinez et al. 2001), phage display techniques (Lowman, Bass et al. 1991) or site directed mutagenesis. Understanding of sequence to structure correlations and their implications for protein function transcends specific protein structures and functions. The significance of quality experimental data for computational efforts in developing protein structure heuristics from sequence to structure correlations cannot be overstated. The evolutionary correlation based contact prediction method achieves on average comparable accuracy to related methods and novel features in predicted protein contact maps. The input data preparation and prediction algorithm development and improvement represent a novel approach in the area of protein structure prediction. The resulting computational application should shed light on protein structure prediction and evolutionary correlation heuristics.

Methods

MSA, Phylogenetic and Structural Data

For survey and test cases of contact prediction, sequences of known protein structure were used as queries for the PSIBLAST algorithm (Altschul, Madden et al. 1997) against the GenPept database from the National Center for Biotechnology Information. A minimum of sequence information relating to evolutionary diversity was required, and families with less than 15 sequences were rejected. Sequences were aligned and phylogenetic trees created with CLUSTALW (Thompson, Higgins et al. 1994) and/or combinations of software from the GCG package (Devereux, Haeberli et al. 1984) including PILEUP and PAUPSEARCH (Rogers and Swofford 1999).

Definition of Protein Amino Acid Contacts

Based on known protein crystal structures, distances between the Cβ atoms of amino acid side chains, or the Cα atom of glycine, were used to compute inter-residue distances. Residue contacts were determined by a fixed 8 Å

distance cutoff, with the additional criteria of no other atoms in between the two contacting residues. An additional criterion of no atoms in between the contacting residue pair improves the accuracy of the residue-residue contact model (W.R. Taylor, personal communication). The resulting binary contact information was encoded in an *n* by *n* matrix, *n* being the sequence length of the protein structure or MSA.

Definition of Evolutionary Correlation

Based on the ideas of the ET method (Lichtarge, Bourne et al. 1996), we used a definition of sequence invariance within subclades of a phylogeny (Figure 2 and 3). Evolutionary correlations were calculated directly from the phylogenetic and multiple sequence data. The evolutionary correlation score was represented by the sum of invariance differences (see Figure 3 and Chapter III), incremented for every case of pairwise subclade invariance of pairs of positions.

Calculation of Evolutionary Sequence Correlations

The initial data for prediction of protein contact maps was generated based on the underlying sequence alignment of sequences in the evolutionary family. Conserved positions and positions with greater than 30 % gaps within the family, were excluded in the final algorithm. The concept of evolutionary correlation was defined as the invariance of pairs of positions in different subclades, and of different amino acids. See Chapter III for a detailed discussion of subclade invariance in phylogenetic data. This pairwise correlation definition was applied to the MSA data across all possible pairwise combinations of subclades in the phylogeny. The overall evolutionary correlation score for a pair of positions x and y **ECS(x,y)** was calculated as the following sum:

$$\mathbf{ECS(x,y)} = \sum \mathbf{A}_{ij}$$

where A_{ij} = 1 if positions x and y are invariant in subclades i,j and A_{ij} = 0 if positions x and y are not invariant in subclades i,j

where i and j are subclades of the phylogeny. The sum was taken over all the subclade pairs within the protein family, limited and normalized using intervals of sequence distances. This procedure resulted in evolutionary scores for pairs of positions in the multiple alignment (Figure 3). Subsequently, additional parameters were incorporated into the evolutionary correlation scoring function, notably the difference of the percent sequence identities of the pair of subclades. This modification was used to favor divergent correlations (see Figure 2). Correlations derived from divergent subclades contained fewer false positive contact correlations, improving the signal to noise ratio. Positions identified with invariance comparisons were then used as input for contact map prediction.

Sorting Correlations

In order to more closely reflect the nature of protein structure and protein contact maps, contact predictions were limited to a specific number of top correlations. The number used in the final algorithm was equal to the sequence length of the protein in question. A number of other limits for the number of predicted contacts were tested, however no useful trends were identified aside from

the observation that overall predicting fewer contacts resulted in greater information content improvements. A similar sorting procedure was reported by Olmea & Valencia 1997 (Olmea and Valencia 1997).

Measurement of Absolute and Conditional Amino Acid Contact Frequencies

The contact map was divided into two bands parallel to the diagonal of the contact matrix. The bands correspond to intervals of sequence separation: from 4 to the minimum of 25 or $1/5^{th}$ the length of the protein for local contacts, and from the minimum of 25 or $1/5^{th}$ the length of the protein on for non-local contacts (see Figure 8).

A nonredundant subset of the PDB (Harrison 1996), consisting of 431 structures, was used to derive local per-residue contact frequencies in protein structures. The unconditional local and prior non-local contact occupancies were also calculated using the nonredundant data set. The conditional non-local contact probability was calculated by using the formula of Bayes theorem for the probability of $P(A_i|B)$:

$$\frac{P(B|A_i) \cdot P(A_i)}{\sum P(B|A_i) \cdot P(A_i)}$$

Where $P(A_i|B)$ is the probability of event A_i given event B. The local contact occupancy was assigned as the unconditional distribution $P(A_i)$. The prior distribution $P(B|A_i)$ was approximated by contact distribution end effects, considering conditional probabilities of contact for i&n given unconditional contacts and noncontacts of i. The known structures were analyzed for specific instances of residues with *x* local contacts and *y* non-local contacts. This data was recorded in a 15 by 15 matrix and normalized across all measurements (see Figure 8).

Derivation of a Pairwise, Sequence Separation Based Amino Acid Interaction Potential

A pairwise amino acid interaction potential was derived from a nonredundant set of 431 structures (Harrison 1996). The potential calculation was as described previously (Sippl 1990). The following modifications were implemented: the considered interval of sequence separation was 4 to 42 residues, with the last interval encoding all

contacts with a sequence separation greater than 41. Residue-residue distances were distinguished from 2 Å to 9 Å in 1 Å intervals, and the definition of protein residue contacts used was as above. Importantly, the protein structure data set used was considerably larger than that of Sippl (Sippl 1990). The statistical mechanical equation for free energy was used to convert frequencies of pair residue and contact occurrences at specific sequence separations s was the following:

$$G_s = kT \ln (f_{cs} / f_s)$$

Where f_s corresponds to the occurrence frequency and f_{cs} to the contact frequency of an amino acid pair respectively.

A figure of the frequency of contacts across the analyzed interval of sequence separation is shown (Figure 13). An example of the 20 by 20 interaction potential matrix for sequence separation of 20 is also shown (Figure 14).

Figure 13

Results for sequence separation calculations for all contacts found in a nonredundant subset of the PDB. Pairs of residue-residue contacts with sequence separation of 4 to 40 were considered.

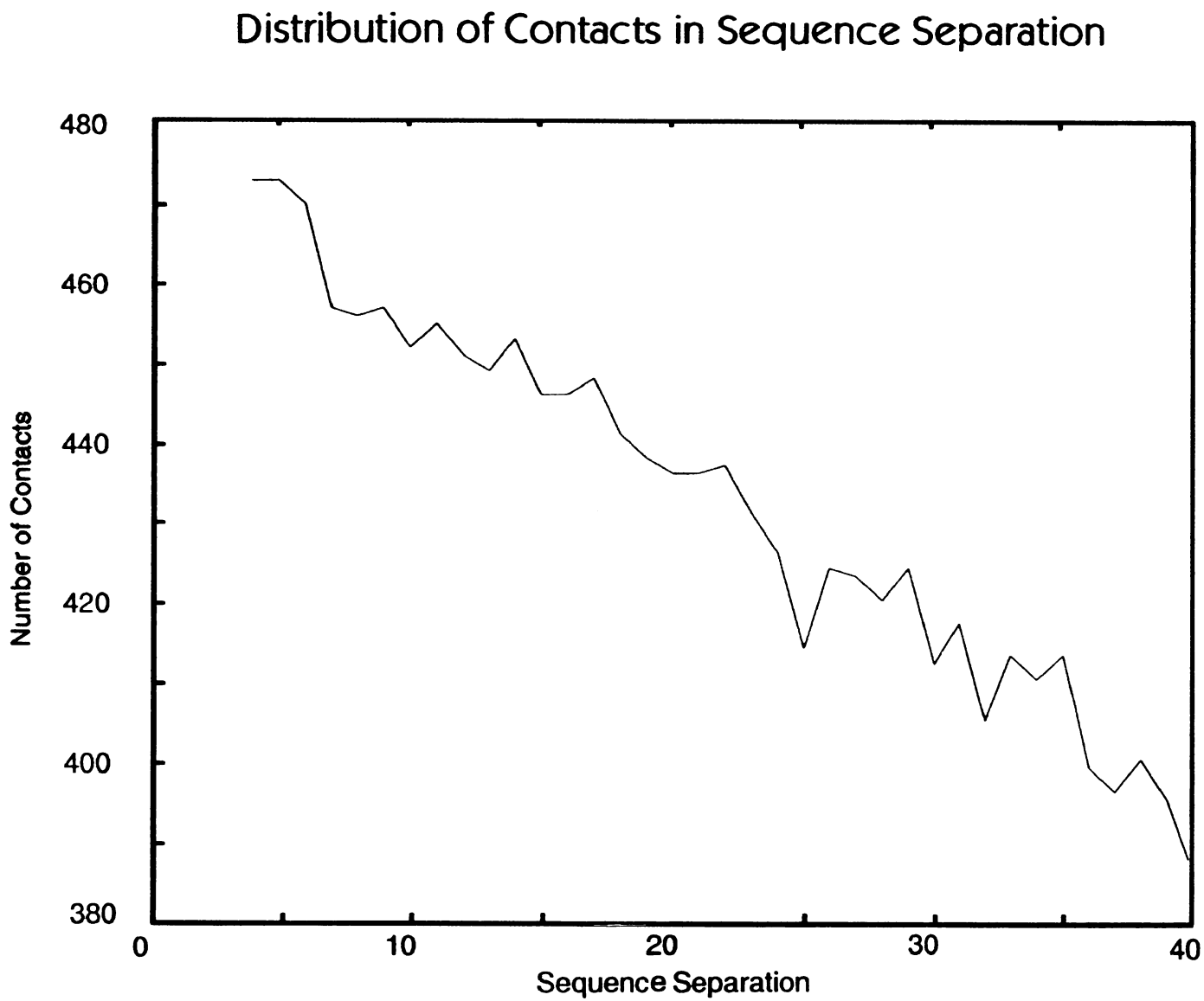

Distribution of Contacts in Sequence Separation
480
460
440
420
400
380
Number of Contacts
0
10
20
30
40
Sequence Separation

Figure 14

Example of an amino acid pair substitution matrix in a sequence separation interval and the approximation of potential energy of amino acid interactions. Frequencies of amino acid pairs were calculated for intervals of residue-residue distances and converted to potential energy (see Methods). The matrix corresponds to a pairwise interaction potential, calculated by normalizing the number of amino acid contacts of a given type, and by the sequence occurrence of the pair at a sequence separation (see Methods). The potential is colored using a blue to red to white color scale, where blue represents favorable and white unfavorable interactions. Example of symmetries and asymmetries in the interaction energies are shown. The potential was calculated by assuming a direction of the chain, hence the upper and lower diagonals of the matrix are not symmetric.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	Y	W
A	0.77	0.83	0.83	0.37	0.83	0.83	0.83	0.37	0.83	0.79	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
C	0.83	0.0	0.83	0.83	0.83	0.66	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.37	0.83	0.83	0.83	0.83	0.83
D	0.83	0.83	0.55	0.83	0.83	0.83	0.83	0.37	0.15	0.83	0.83	0.19	0.83	0.83	-0.0	0.37	0.26	0.66	0.55	0.22
E	0.83	0.83	0.37	0.83	0.83	0.83	0.26	0.83	0.23	0.83	0.83	0.37	0.83	0.55	0.04	0.15	0.83	0.83	0.37	0.02
F	0.37	0.83	0.83	0.83	0.55	0.83	0.83	0.83	0.83	0.30	0.83	0.30	0.83	0.83	0.83	0.55	0.55	0.83	0.83	0.37
G	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.66	0.66	0.83	0.37
H	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.55	0.83	0.19	0.83	0.83	0.19
I	0.83	0.83	0.83	0.83	0.66	0.83	0.83	0.37	0.55	0.48	0.37	0.83	0.83	0.83	0.83	0.55	0.83	0.55	0.83	0.83
K	0.83	0.83	-0.1	0.38	0.83	0.83	0.83	0.83	0.55	0.83	0.83	0.83	0.83	0.37	0.83	0.37	0.55	0.83	0.83	0.19
L	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.55	0.83	0.83	0.83
M	0.55	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.55	0.37	0.83	0.83	0.83	0.37	0.83
N	0.83	0.83	0.83	0.66	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.73	0.83	0.55	0.56	0.19	0.19	0.55	0.26	0.83
P	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.37	0.83	0.83	0.83	0.83	0.83
Q	0.83	0.83	0.30	0.37	0.37	0.51	0.37	0.83	0.37	0.83	0.83	0.83	0.83	0.26	0.37	0.26	0.66	0.83	0.83	0.21
R	0.83	0.83	0.01	-0.1	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.15	0.83	0.19	0.83	0.83	0.37	0.19	0.83	0.13
S	0.83	0.55	0.19	0.[illegible]	0.55	0.83	0.83	0.83	0.83	0.37	0.83	0.83	0.83	0.66	0.08	0.79	0.26	0.83	0.83	0.19
T	0.83	0.83	0.55	0.83	0.83	0.83	0.37	0.83	0.83	0.83	0.83	0.37	0.83	0.55	0.55	0.10	0.18	0.83	0.83	0.83
V	0.83	0.83	0.83	0.83	0.37	0.83	0.83	0.83	0.83	0.66	0.83	0.37	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.37
Y	0.83	0.83	0.83	0.37	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.37	0.83	0.83	0.83	0.83	0.83	0.83
W	0.83	0.83	0.83	0.37	0.83	0.83	0.37	0.37	0.83	0.19	0.83	0.83	0.83	0.83	0.55	0.66	0.66	0.37	0.83	0.37

R to W

E to R

W to R

R to E

Prediction of Protein Secondary Structure

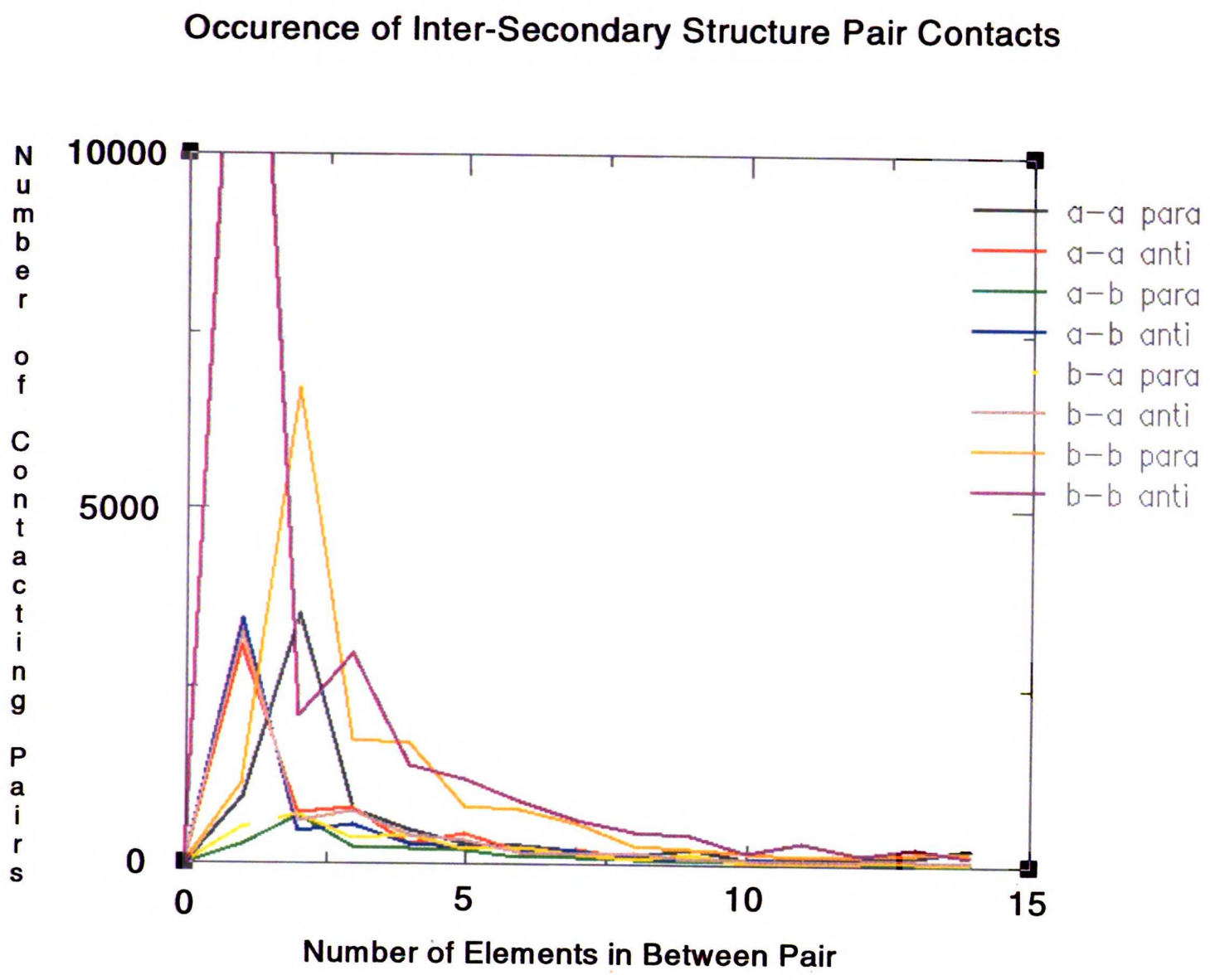
Secondary structure was predicted using the multiple sequence data used for contact prediction with the method of Chandonia & Karplus 1999 (Chandonia and Karplus 1999), using default parameters.

Measurement of Protein Secondary Element Interaction Frequencies

To obtain information on the patterns of tertiary contacts between secondary structure elements, Bayesian statistics were used to calculate conditional contact probabilities for residues in secondary structure elements. Alpha helix-alpha helix, alpha helix-beta strand and beta strand-beta strand secondary structure interactions were distinguished. A survey of the occurrence of each secondary structure interaction type is given in Figure 15. The survey was based on a nonredundant subset of the PDB with 431 structures (Harrison 1996). Additional parameters were the number of secondary structure elements between the two elements (ranging from 1 to 14), and the orientation of the element interaction (classified as parallel or antiparallel). The orientation of the interaction was

Figure 15

A survey of secondary structure element interaction types in protein structures. A nonredundant subset of the PDB database was used to calculate frequencies of specific interactions (see Methods). Antiparallel beta-sheets are the most abundant, while parallel beta-sheet interactions are second in occurrence. Most inter-secondary structure element contacts occur between elements separated by one to four secondary structures.



determined by calculating the dot product of the two vectors defined by the Cβ (Cα if glycine) atoms of the first and last residues of the each secondary structure element. A product of greater than zero corresponded to a parallel orientation, and a product of less than zero corresponded to an antiparallel orientation.

Only secondary elements in contact, i.e. with at least one pair of residues satisfying the residue-residue contact definition, were used in the derivation (Figure 15). Contact probabilities were encoded in 21 by 21 matrices, were the center of the matrix (10,10), corresponded to the center of interaction of the two secondary structure elements. This matrix was used as a representation of the contact occupancy space for a pair of secondary structure elements (Figure 16). The element pair center was identified by considering the middle residue of each element and assigning the middle residues to the center of the contact probability matrix. All contacts between the two elements were then recorded in the matrix using the element pair center as a reference to align the contact spaces of the two secondary structures (Figure 16). An example of the resulting secondary structure contact probability matrix for alpha-alpha, beta-beta and alpha-beta is shown (Figure 17).

Figure 16

Diagram of the procedure used to calculate and apply secondary structure interaction probabilities. Predicted or actual protein secondary structure was used to assign an interaction space between two elements (see Methods). The space is represented as a square matrix. The center of the matrix (red square) corresponds to the pair of middle residues of the two elements (see Methods). This center was used as a reference point to evaluate or predict residue contacts between the two elements.

Reconstruction of Inter-Secondary Structure Element Residue Contacts

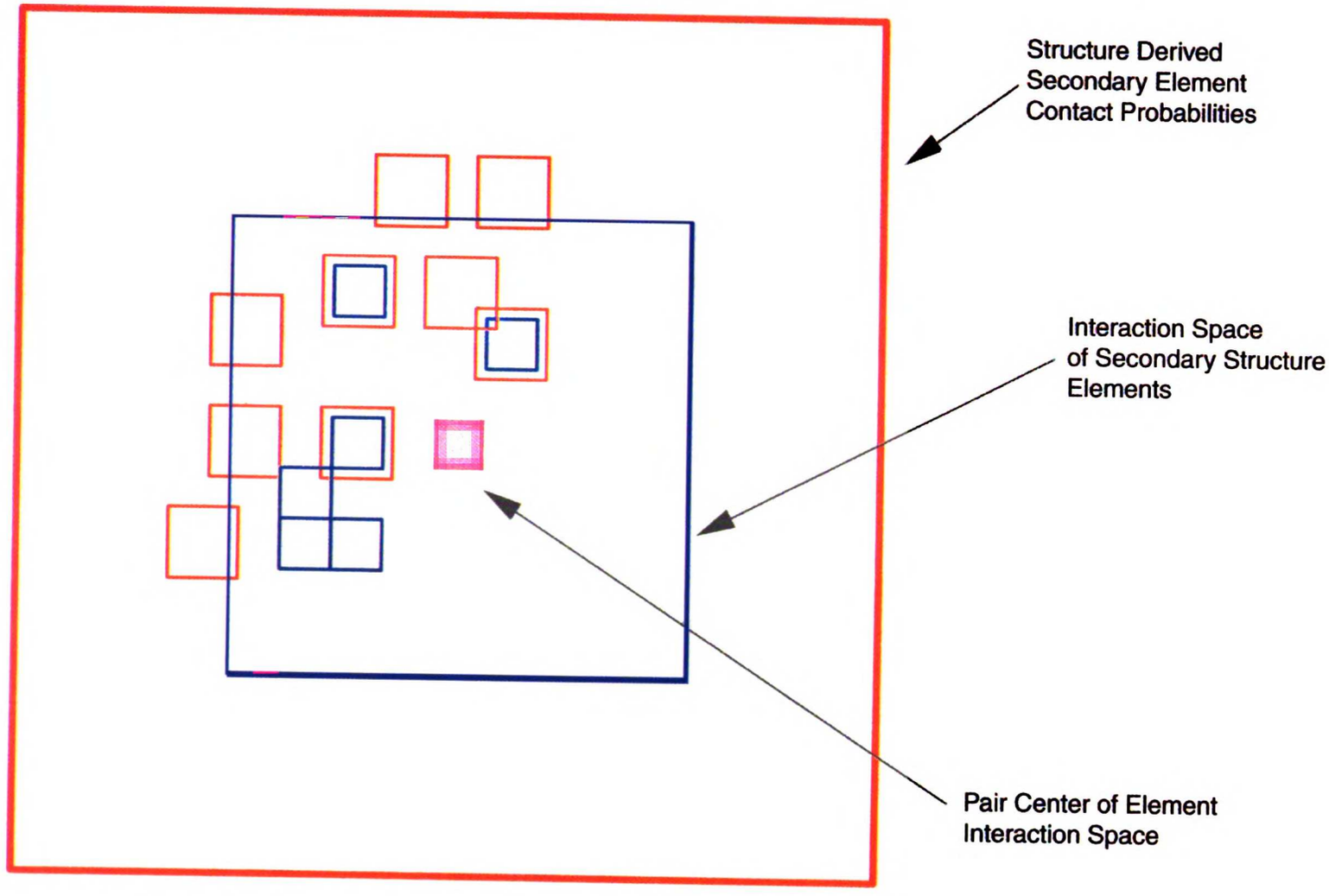
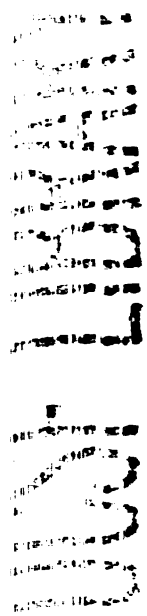
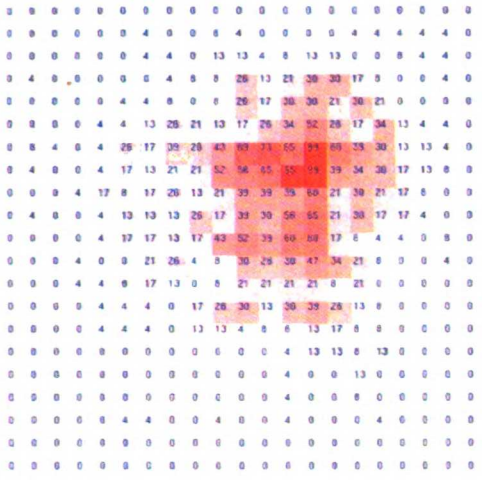


Figure 17

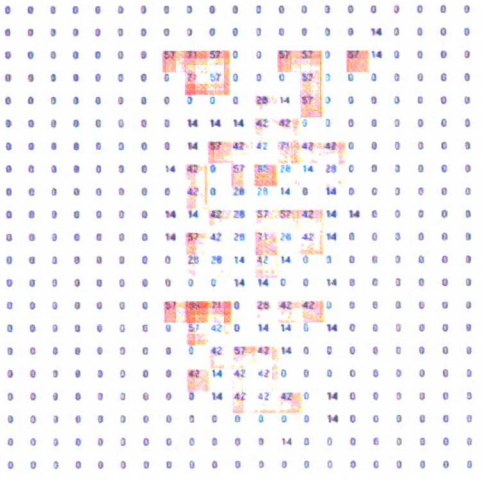
Examples of matrices representing the calculated unconditional secondary structure contact probabilities (see Methods). The label for each matrix describes the contact types distinguished in the calculation. These are the number of secondary structure elements between the two elements in consideration, the relative orientation of the interaction and the types of secondary structures involved. For example, the label "1 : parallel : alpha-alpha" corresponds to a parallels interactions between a pair of alpha helices separated by one other secondary structure element.



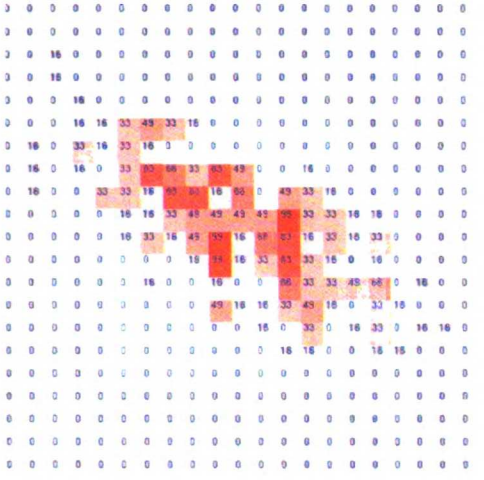
1 : parallel : alpha-alpha



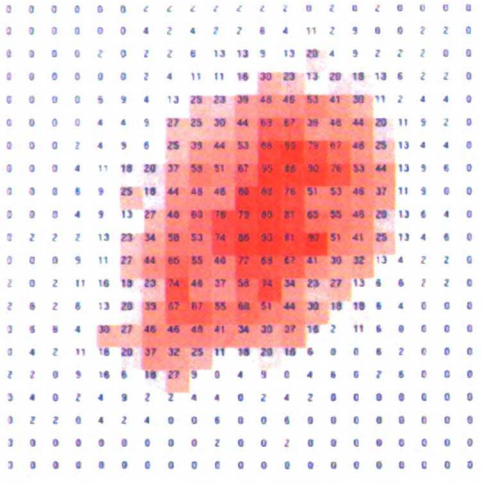
5 : parallel : beta-alpha



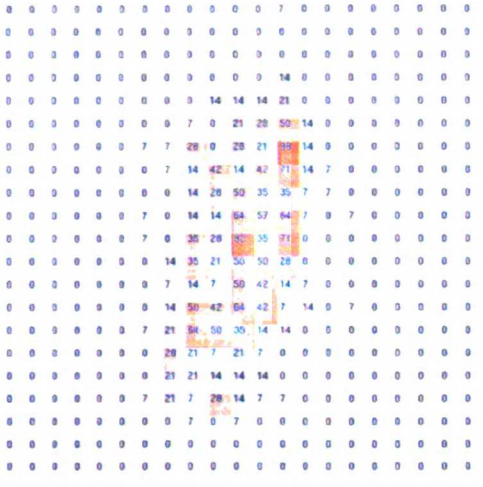
5 : parallel : alpha-beta



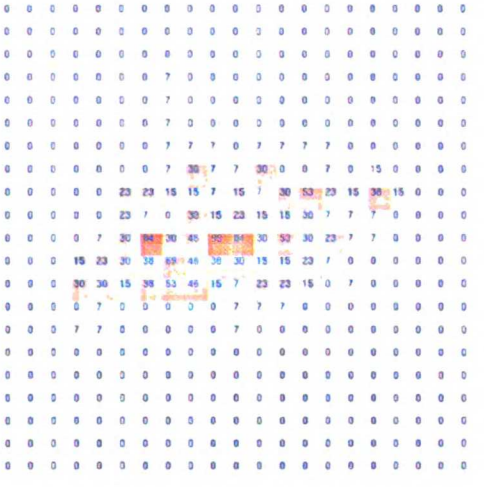
1 : antiparallel: alpha-alpha



5 : antiparallel: beta-alpha



5 : antiparallel: alpha-beta



Bayesian Statistics of Protein Secondary Element Interaction Frequencies

The element pair center contact frequencies described above, were used as the unconditional distribution, $P(A_i)$, for calculation of Bayesian statistics of the secondary structure interactions. To approximate the prior distribution $P(B|A_i)$, end effects in distributions of contacts between secondary structure were calculated by considering the i±n and j±n contacts for both i,j contacts and noncontacts. The calculation was based on a nonredundant subset of the PDB with 431 structures (Harrison 1996). Conditional probabilities were calculated based on specific conditions of secondary interactions, e.g. given an antiparallel interaction between an alpha helix and a beta strand separated by 4 other secondary structures (Figure 18).

The Bayes theorem formula was used to calculate the conditional probability. The conditional probability of event A_i given event B, $P(A_i|B)$, is given by:

$$\frac{P(B|\ A_i)\cdot P(A_i)}{\Sigma\ P(B|A_i)\cdot P(A_i)}$$

Figure 18

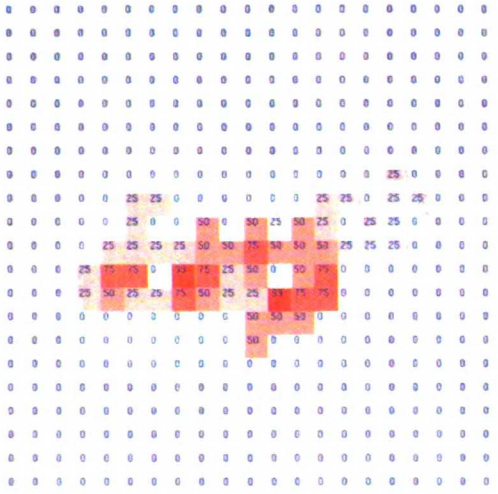
Examples of matrices representing the calculated conditional secondary structure interaction probabilities (see Methods). The designation for each matrix describes the contact types distinguished in the calculation. These are the number of secondary structure elements between the pair of elements under consideration, the relative orientation of the interaction and the types of secondary structures involved. The first line of the designation describes the condition of the interaction, while the second the conditional interaction. For example, the designation:

1 : parallel : alpha-alpha

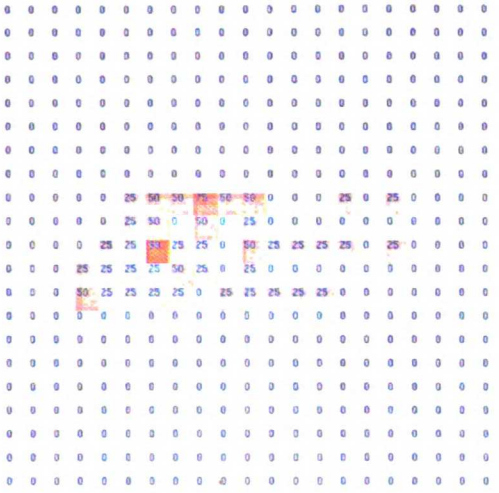
5: antiparallel: alpha-beta

corresponds to an interaction condition of a parallel interaction between a pair of helices separated by one other secondary structure element, for the conditional antiparallel interaction of an alpha helix with a beta strand separated by 5 other secondary structure elements.

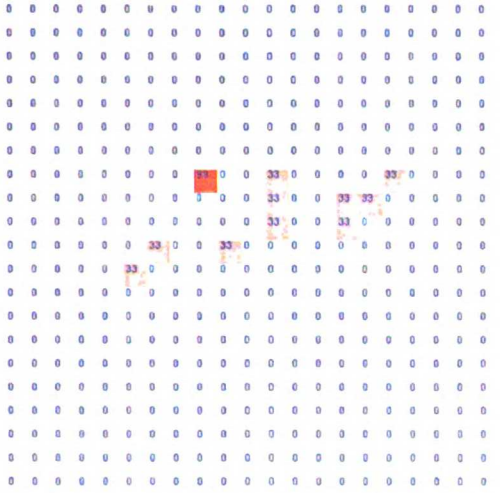
1 : parallel : alpha-alpha
5: antiparallel: alpha-beta



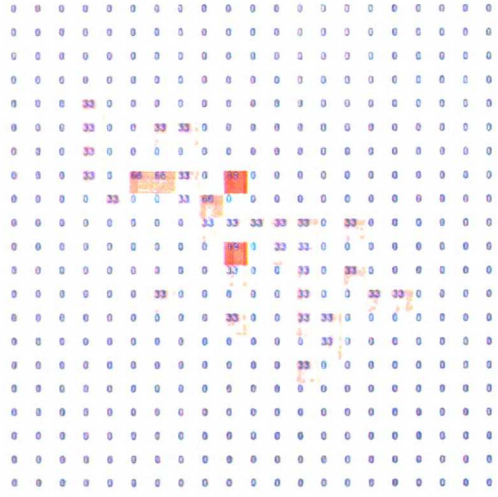
5 : parallel : alpha-beta
5: antiparallel: alpha-beta



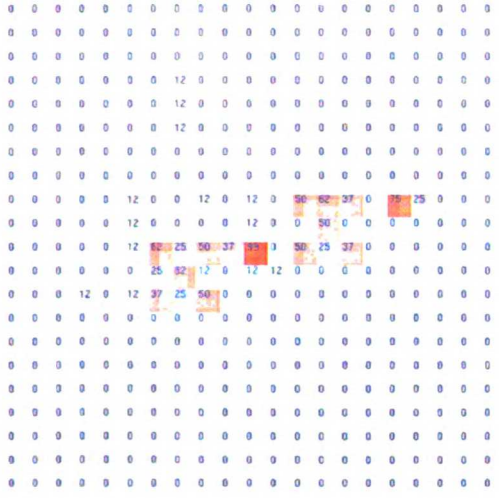
5 : parallel : beta-alpha
5: antiparallel: alpha-beta



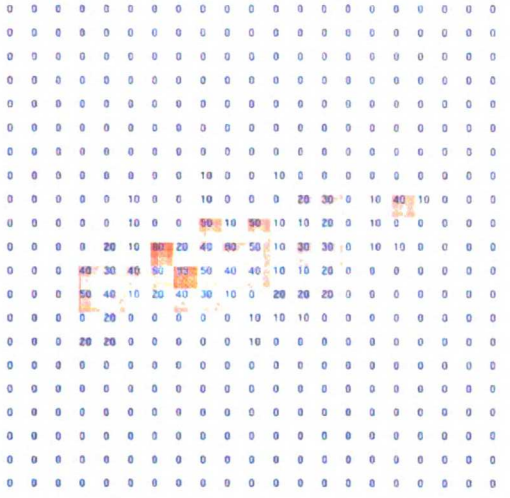
5 : antiparallel : alpha-beta
5: parallel: alpha-beta



5 : antiparallel : alpha-beta
5: antiparallel: alpha-beta



5 : antiparallel : beta-alpha
5: antiparallel: alpha-beta



A_i represents a specific secondary structure interaction condition, and B represents the probability of a secondary structure interaction conditional on A_i.

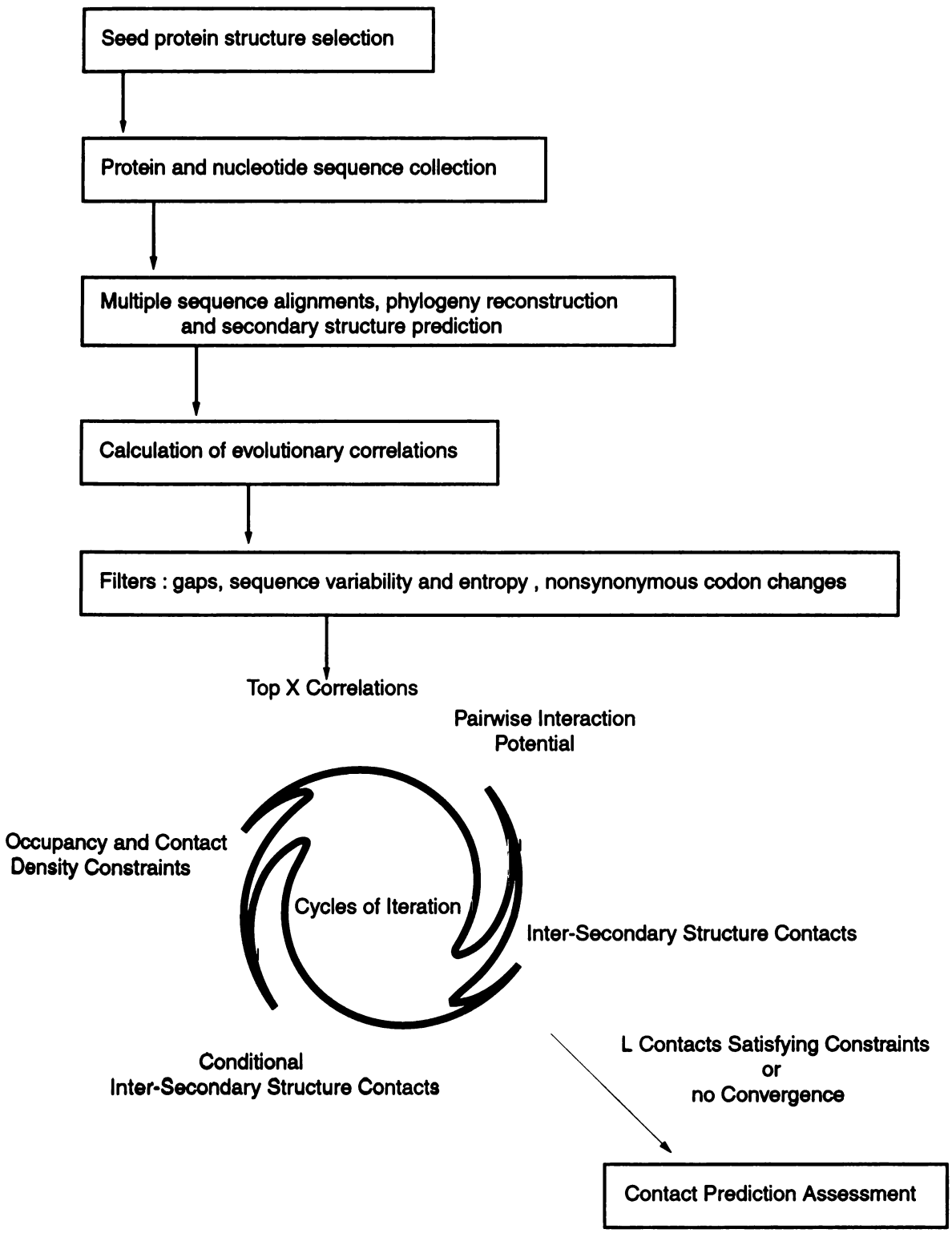
Iterative Algorithm for Contact Map Prediction

Figure 19 depicts the flow diagram of the iterative algorithm used to predict protein tertiary structure amino acid contacts.

The discrete steps and multiple iterations of the algorithm were programmed in JAVA. The resulting application included methods used to process and manipulate the input data, relying on many of the same objects and methods as the algorithm itself. To drive the iterations, a number of structural conditions were analyzed at each step of the contact map prediction. These measures included, contact coordination, neighborhood contact occupancy, the predicted number of contacts in intervals of sequence separation and conditional measures of local, non-local and inter-secondary element contacts. The average of the absolute difference between the database derived local contact frequencies and the predicted local contact frequencies was used to determine the local density of predicted contacts.

Figure 19

A flow diagram of the prediction procedure implemented in the algorithm. After initial data collection and manipulation, the evolutionary correlations were calculated from the input data. A number of filters were applied to the raw evolutionary correlations. Contact prediction proceeded iteratively by application of secondary structure interaction probabilities and acceptance or rejection of a contact prediction based on contact occupancy constraints and conditional probabilities.


Seed protein structure selection
Protein and nucleotide sequence collection
Multiple sequence alignments, phylogeny reconstruction and secondary structure prediction
Calculation of evolutionary correlations
Filters : gaps, sequence variability and entropy , nonsynonymous codon changes
Top X Correlations
Pairwise Interaction Potential
Occupancy and Contact Density Constraints
Cycles of Iteration
Inter-Secondary Structure Contacts
Conditional Inter-Secondary Structure Contacts
L Contacts Satisfying Constraints or no Convergence
Contact Prediction Assessment

Outcomes of these analyses were used to formulate rules for accepting or rejecting predicted interactions or to finalize the contact map prediction. The method was fully automated, and the final predictions encompassed 26 protein families with representative crystal structures.

Contact Prediction Assessment

Three measures of prediction success were used to assess the contact predictions. All measures required presence of a known crystal structure.

A simple signal to noise measure was implemented to assay the information content per protein structure prediction. This measure corresponded to the ratio of correct predicted contacts to all predicted contacts, a more standard concepts of prediction accuracy.

The Matthews correlation coefficient (Matthews 1975), defined by:

$$\frac{C_1 * C_2 + C_3 * C_4}{((C_2 + C_3) * (C_2 + C_4)(C_1 + C_3)(C_1 + C_4))^{1/2}}$$

Where C_1 are correct predicted contacts, C_2 are correct predicted noncontacts, C_3 are underpredicted true contacts,

C_4 are overpredicted true noncontacts. C_3 and C_4 are especially useful since they capture both underpredicted as well as overpredicted information.

The final measure of prediction accuracy used was the improvement over a random prediction, defined by:

$$\frac{\dfrac{\text{Predicted correct contacts}}{\text{Total predicted contacts}}}{\dfrac{\text{Real contacts}}{\text{All Possible contacts}}}$$

The numerator of this measure corresponds to the simple signal to noise measure. The original use of an actual structure in predicted contact normalization was reported by Thomas et al 1996 (Thomas, Casari et al. 1996), who applied the denominator to normalize all contacts and noncontacts for length dependent contact map density. This signal to noise measure, where the real protein structure is chosen as a random model, allows the comparison of the signal to noise ratios in the predicted and actual contact data. A similar procedure was later adopted by Fariselli et al (Fariselli, Olmea et al. 2001).

JAVA application

A JAVA algorithm for input data manipulation and contact prediction accompanied by a graphical interface was constructed. WTREETO takes as input a protein phylogeny, MSA and protein structural data. Among its functions are searching for invariant positions and methods for analyzing phylogenetic correlations. WTREETO implements the methods described in the above Methods sections, including the iterative procedure of protein contact map prediction.

WTREETO derives properties of phylogenies in relation to the sequence and structure of family members. Moreover it allows translation of results among these three biological dimensions. Alignment errors, often made clear when analyzing sequence and structure, can be immediately remedied by alignment editing features.

Acknowledgements

Deep and sincere thanks are due to my advisor Fred E. Cohen, as well as Dietlind Gerloff, Jonathan Blake, John-Marc Chandonia, Dirk Walther, prof. Tack Kuntz, prof. Patsy Babbit, Andrew Wallace, Olivier Lichtarge, and all UCSF Biophysics students.

Chapter II

The impact of whole genome sequence data on drug discovery - a malaria case study.

Abstract

Background

Identification and validation of a drug discovery target is a prominent step in drug development. In the post-genomic era it is possible to reevaluate the association of a gene with a specific biological function to see if a homologous gene can subsume this role. This concept has special relevance to drug discovery in human infectious diseases, like malaria. A trophozoite cysteine protease (falcipain-1) from the papain family, thought to be responsible for the degradation of erythrocyte hemoglobin, has been considered a promising target for drug discovery efforts owing to the antimalarial activity of peptide based covalent cysteine protease inhibitors. This led to the development of non-peptidic non-covalent inhibitors of falcipain-1 and their characterization as antimalarials. It is now clear from sequencing efforts that the malaria genome contains more than one cysteine protease and that falcipain-1 is not the most important contributor to hemoglobin degradation. Rather, falcipain-2 and

falcipain-3 appear to account for the majority of cysteine hemoglobinase activity in the plasmodium trophozoite.

Materials and Methods

We have modeled the falcipain-2 cysteine protease from one of the major human malaria species, *Plasmodium falciparum* and compared it to our original work on falcipain-1. As with falcipain-1, computational screening of the falcipain-2 active site was conducted using DOCK. Using structural superpositions within the protease family and evolutionary analysis of substrate specificity sites, we focused on the commonalities and the protein specific features to direct our drug discovery effort.

Results

Since 1993, the size of the Available Chemicals Directory had increased from 55313 to 195419 unique chemical structures. For falcipain-2, eight inhibitors were identified with IC_{50}'s against the enzyme between 1 and 7 μM. Application of three of these inhibitors to infected

erythrocytes cured malaria in culture, but parasite death did not correlate with food vacuole abnormalities associated with the activity of mechanistic inhibitors of cysteine proteases like the epoxide E64.

Conclusions

Using plasmodial falcipain proteases, we show how a protein family perspective can influence target discovery and inhibitor design. We suspect that parallel drug discovery programs where a family of targets is considered, rather than serial programs built on a single therapeutic focus, will become the dominant industrial paradigm. Economies of scale in assay development and in compound synthesis are expected owing to the functional and structural features of individual family members. One of the remaining challenges in post-genomic drug discovery is that inhibitors of one target are likely to show some activity against other family members. This lack of specificity may lead to difficulties in functional assignments and target validation as well as a complex side effect profile.

Introduction

Cysteine proteases play a number of degradative and regulatory roles in a wide range of organisms. One measure of the success of this enzymatic motif is the degree of cysteine protease speciation. For example, the malaria genome is predicted to contain at least 5 cysteine proteases. This protein family is defined by a unique fold, which has speciated functionally many times producing subfamilies with unique substrate specificities. This proliferation of proteases creates the likelihood that more than one enzyme could subsume the same function *in vivo* and complicates the task of identifying the best targets for drug discovery. In the pre-genomic era, drug discovery targets were identified via a reductionist approach where genes were sought that carried out a physiological role. Further proof of principle was obtained using chemical inhibitors of the gene products. If the chemical inhibitor used had a broad specificity, the conclusions reached could be subject to question. In the post-genomic era, the genetic "deck of cards" is known and process of elimination logic can play a more prominent role in the identification of targets.

Given the success of angiotensin-converting enzyme (ACE) inhibitors in the treatment of hypertension (Mark and Davis 2000) and HIV protease inhibitors in AIDS (Tebas and Powderly 2000), proteases have become popular drug discovery targets. However, several protease targets, such as the renin aspartyl protease for hypertension and matrix metalloproteases for cancer and arthritis, have not led to marketable products. These difficulties originated not from problems in the sequencing, cloning or annotation efforts but rather because of the redundant and homeostatic nature of biological systems, including the presence of genes performing back-up functions. The proteolytic cascade of the Renin Angiotensin Aldosterone (RAA) system mediates cleavage of angiotensinogen to angiotensin I by the aspartyl protease renin and subsequent cleavage of angiotensin I to the effector peptide angiotensin II by ACE. By the mid 1990's renin inhibitors were widely known to have negligible effects on hypertension (Fisher and Hollenberg 2001), while to date dozens of ACE inhibitors have been proven to be effective human therapeutics for hypertension in spite of their side effects profile. Renin, the upstream enzyme in this pathway has a single unique substrate. While this molecular specificity would be expected to yield a better target for drug discovery

efforts, compensatory homeostatic mechanisms undermine this thesis.

Like renin in humans, the plasmodial cysteine proteases that degrade hemoglobin exist as a family of homologs in the *P. falciparum* genome. As hemoglobin is the major nutritional source for the parasite in the erythrocytic stage, and proteases have been the target of successful drug discovery efforts, inhibitors of hemoglobin degradation have been sought as a new class of antimalarials. In 1987, Rosenthal et al. identified three *P. falciparum* proteases by gel electrophoresis. Two of these had an active site cysteine (Rosenthal, Kim et al. 1987; Rosenthal and Nelson 1992). A papain-like cysteine protease thought to be necessary for hemoglobin degradation in the trophozoite stage of the malaria human life cycle, and now known as falcipain-1, was cloned and sequenced (Rosenthal and Nelson 1992). In 1993, a model of falcipain-1 based on its sequence homology to papain and actinidin was used in a structure-based drug discovery effort to identify a symmetric acyl-hydrazide inhibitor with antimalarial properties at a 6 μM concentration (Ring, Sun et al. 1993). However, optimization of the lead compound was complicated by difficulties in reconciling the activity of the lead analogs with the model protease structure. In

the past year, *P. falciparum* genomic sequencing efforts led to the identification of a number of homologs of falcipain-1 and it now seems likely that the falcipain-2 and falcipain-3 gene products are the major plasmodial cysteine hemoglobinases (Shenai, Sijwali et al. 2000).

During the erythrocytic phase of the life cycle, malaria parasites rely on hemoglobin degradation as the predominant source of amino acids. Interruption of hemoglobin degradation with mechanistic inhibitors of cysteine protease leads to accumulation of undigested hemoglobin, swelling of the food vacuole and parasite death (Rosenthal, McKerrow et al. 1988). The precise order of events in the hemoglobin degradation pathway still remains to be clarified. In 1994, two aspartyl proteases, plasmepsins I (Francis, Gluzman et al. 1996) and II (Hill, Tyas et al. 1994), were isolated from the *P. falciparum* food vacuole and shown to perform the first cleavage of hemoglobin. Recently plasmepsin II has been shown to cleave other erythrocyte proteins (Le Bonniec, Deregnaucourt et al. 1999). Falcilysin, a plasmodial metallopeptidase, was reported to act against partially degraded hemoglobin fragments (Eggleson, Duffin et al. 1999). However, our understanding of the pathway of hemoglobin hydrolysis remains limited, as falcipain-2 and falcipain-3 also

readily hydrolyze native hemoglobin, while multiple plasmodial aspartic protease genes are predicted from the genome.

The antimalarial properties of peptide fluoromethyl-ketones and vinyl-sulphones as cysteine protease inhibitors (Rosenthal, Wollish et al. 1991; Rosenthal, Olson et al. 1996) have encouraged their evaluation in animal models of infection. Unfortunately, activity against murine malaria required high doses and the toxicity of peptide fluoromethyl-ketones in experimental animals has stalled their development (Rosenthal, Olson et al. 1996). These results amplify our need to understand which proteases are most essential to hemoglobin degradation. Using modeling, drug design and inhibitor studies for the falcipain hemoglobinases, we illustrate how a gene family approach to drug targets can enhance the understanding of biological phenotype and its inhibition, and hence expedite the drug development process.

RESULTS

Comparative analysis of the falcipain-2 and falcipain-1 model structures

The original drug discovery effort directed at falcipain-1 led to structure-activity relationships that could not be reconciled with the protease model structure (Li, Chen et al. 1994; Li, Chen et al. 1996). Analogs of the acyl-hydrazide lead compound designed to take advantage of specific interactions in the protein's binding sites were synthesized. Unfortunately, these customized analogs did not lead to improvements in inhibitor affinity (Li, Chen et al. 1994; Li, Chen et al. 1996). In retrospect, we attribute this to the fact that the falcipain used in these assays was purified from parasite extract that is now known to be predominantly falcipain-2. Thus, the design was directed against falcipain-1 but the compounds were tested against predominantly falcipain-2. Inhibition studies are now conducted with recombinant falcipain-2.

Falcipain-2 and falcipain-1 share 37 % sequence identity in the mature protease domain. The original model of falcipain-1 was based on papain and actinidin crystal

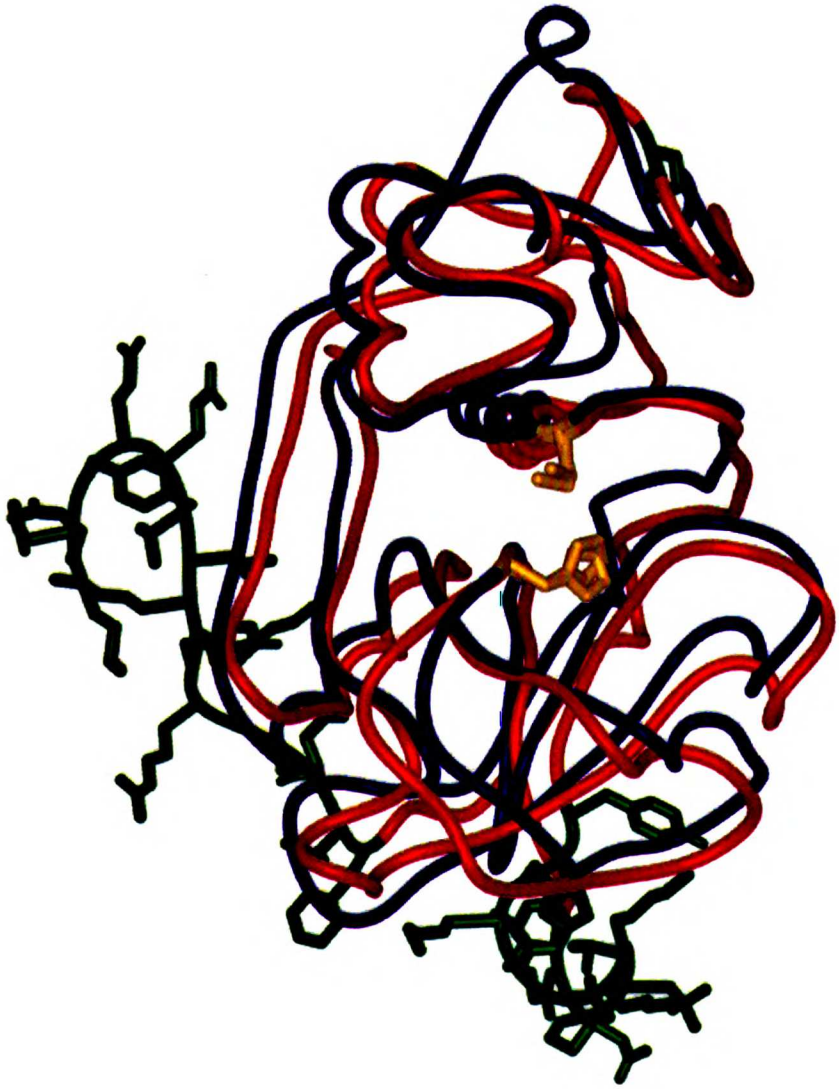
structures (Ring, Sun et al. 1993), and both of the templates were 33% identical to falcipain-1 in sequence. The current model of falcipain-2 is based on a crystal structure of human cathepsin K, which is 35 % identical to falipain-2 in sequence over the mature protease domain. Using standard homology modeling procedures, we constructed a model of the active form of the falcipain-2 hemoglobinase (see Methods).

A comparison of the falcipain-1 and falcipain-2 models was carried out to determine features that could explain functional differences of these proteases and to direct our drug design efforts. The majority of the sequence changes, and all of the three insertion-deletion events, are on the face of the protein opposite the active site (see Figure 20). Nevertheless, there are a number of changes that significantly alter the features of the specificity sites, predominantly on the non-prime side of the peptide binding cleft that recognizes the side-chains on residues N-terminal to the cleavage site.

Protease specificity is frequently studied in the context of subsites that flank the catalytic residues and provide the enzyme with specific preferences for peptide or protein substrates. Following the nomenclature of Berger and Schechter (Berger and Schechter 1970), these sites are

Figure 20

Superposition of falcipain-1 and falcipain-2 model structures. A structural alignment of the falcipain-2 (red) and falcipain-1 (blue) model structures was performed with MINAREA (Falicov and Cohen 1996). Structural positions present in falcipain-2 and absent in falcipain-1 are colored green. The catalytic dyad is shown in yellow for reference. The figure was generated with CHIMERA (Huang, Couch et al. 1996).



referred to as S_4, S_3, S_2, S_1, S_1', S_2', S_3' (see Figure 21), and they correspond to substrates with the sequence P_4, P_3, P_2, P_1, P_1', P_2', P_3', where the P_1-P_1' peptide bond is cleaved. The papain cysteine protease family has well-defined sites from S_3 to S_1', with some individual proteases having more extended specificity. The S_2 and S_1 sites contribute the strongest preference to substrate binding in the case of falcipain-2 (Shenai, Sijwali et al. 2000) and many other papain family proteases (McGrath 1999).

In all, there are 11 amino acid differences in the S_2, S_3 and S_4 sites of falcipain-2 relative to falcipain-1 (see Figure 22). The most variable S_2 site has a total of six differences ranging from conservative to functionally significant ones (see Figure 22). The combination of conservative changes retains the overall hydrophobic character of this site; however, there is a net gain of two non-hydrogen atoms in side-chains on the part of falcipain-1, decreasing the free volume available for binding in this site. Two pairs of these sequence differences appear to be compensating substitutions: S46A and A175S conserve the serine, while N86F and S149N conserve the asparagine (see Figure 22). At the other end of the spectrum, the sequence difference I85P is predicted to have a pronounced effect on the local backbone geometry. This is evident from a

Figure 21

The falcipain-2 model specificity sites. The residues that line the substrate specificity sites are displayed on the falcipain-2 model. The natural peptide substrate orientation prefers the non-prime side for the *N*-terminus, and the prime side for the *C*-terminus. Residues at the boundaries of a site may contribute to neighboring sites. The figure was generated with WEBMOL (Walther 1997) and a sequence to structure alignment and family analysis JAVA™ package (Chapter III).

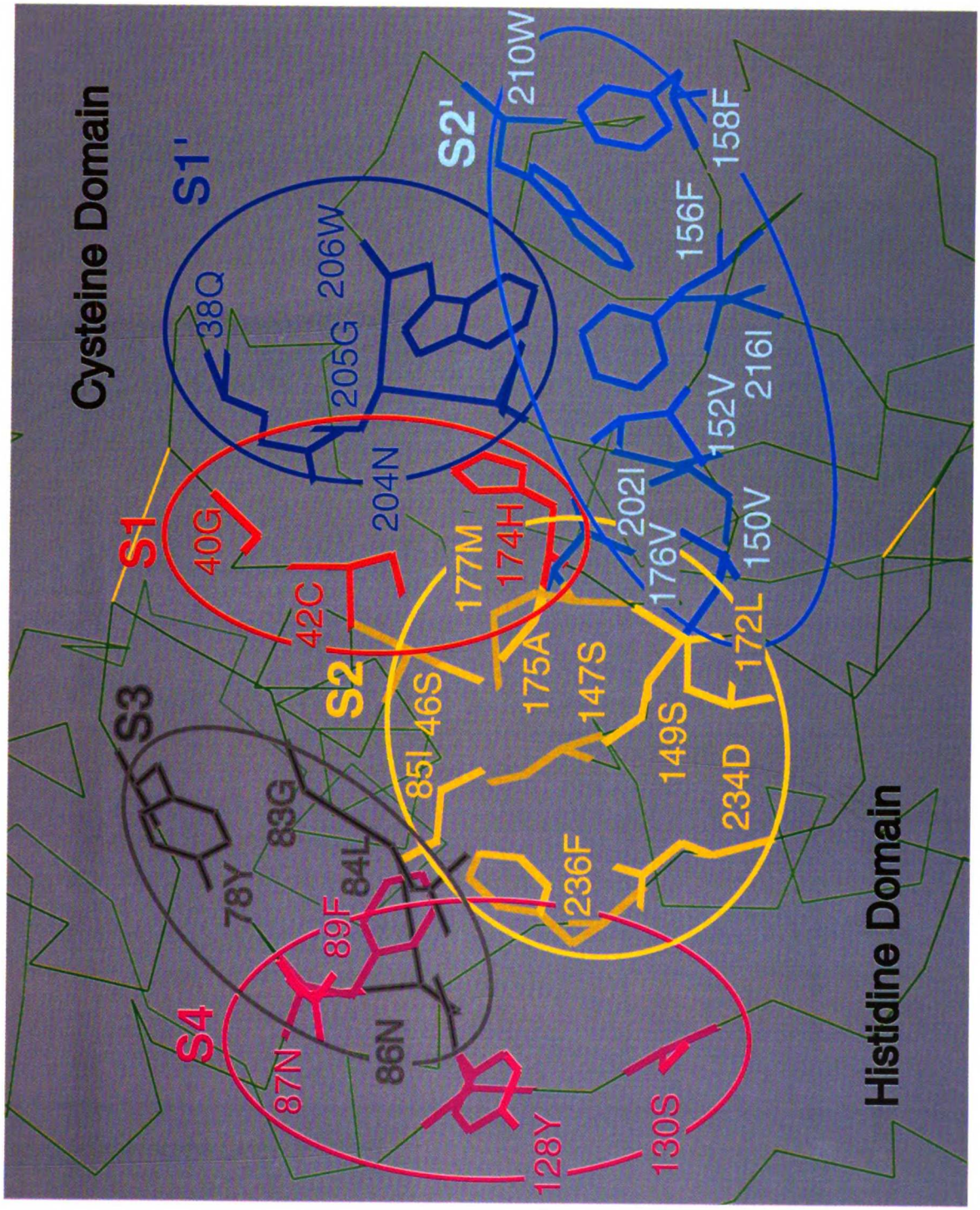
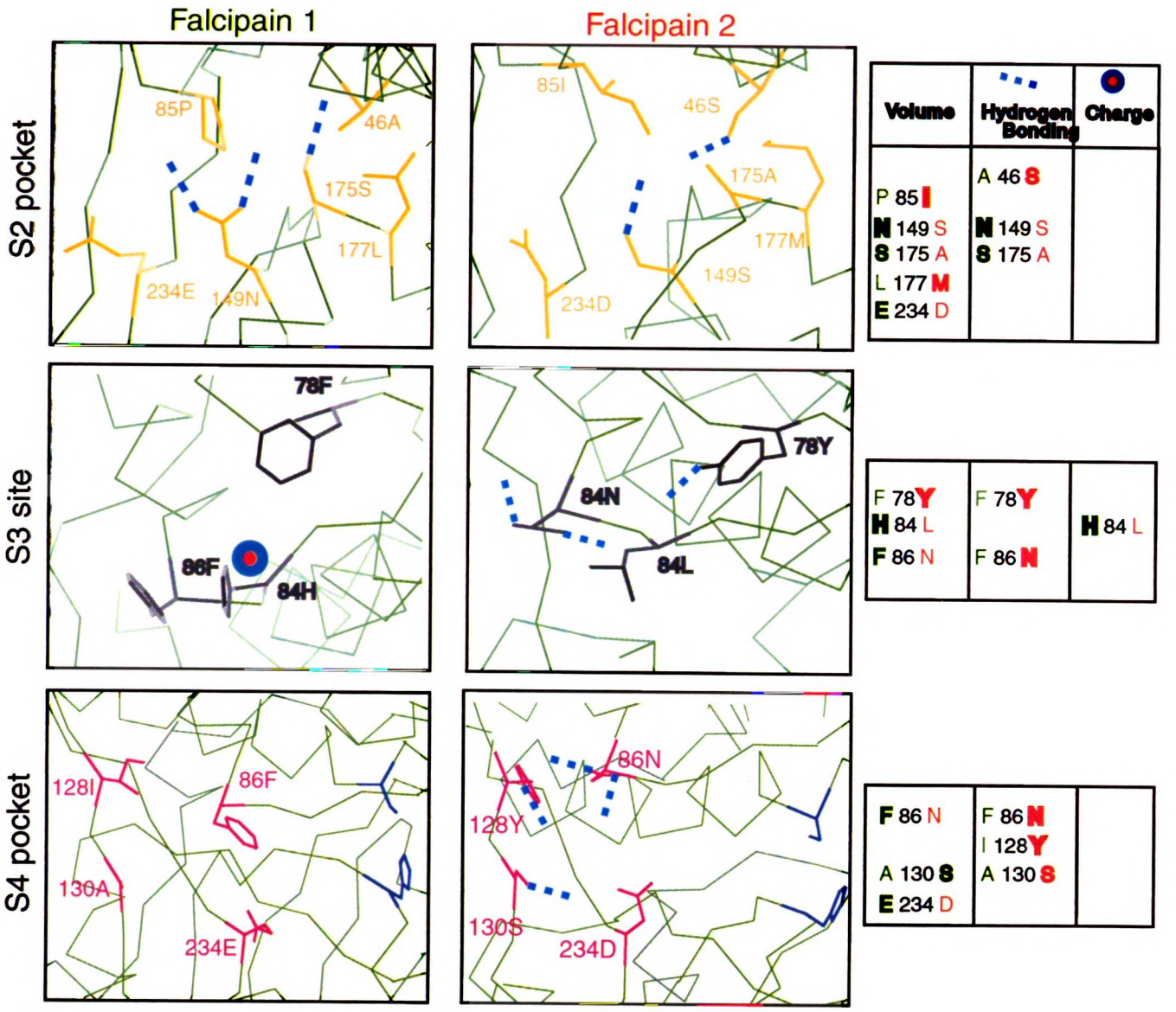

Cysteine Domain
S1'
38Q
206W
205G
204N
S2'
210W
156F
158F
152V
216I
202I
176V
150V
172L
S1
40G
42C
S2
177M
174H
175A
147S
46S
85I
149S
234D
236F
S3
78Y
83G
84L
89F
86N
S4
87N
128Y
130S
Histidine Domain

Figure 22

Falcipain-1 versus falcipain-2 S_2, S_3 and S_4 specificity site analysis. To highlight differences that affect the S_2, S_3 and S_4 specificity sites, the sites were analyzed with respect to sequence differences between the two plasmodial proteases (see Methods). Residues that gained hydrogen bonding functionality relative to the other sequence are marked with blue dashes, residues that gained charge functionality are marked by a red ball within a blue ball. The table of residue changes highlights the sequence that gained functionality with a boldfaced font. The figure was generated with WEBMOL (Walther 1997) and a sequence to structure alignment and family analysis JAVA™ package (Chapter III).


Falcipain 1
Falcipain 2
S2 pocket
S3 site
S4 pocket
85P
46A
175S
177L
234E
149N
85I
46S
175A
177M
149S
234D
78F
86F
84H
78Y
84N
84L
128I
86F
130A
234E
86N
128Y
130S
234D
Volume
Hydrogen Bonding
Charge
P 85 I
N 149 S
S 175 A
L 177 M
E 234 D
A 46 S
N 149 S
S 175 A
F 78 Y
H 84 L
F 86 N
F 78 Y
F 86 N
H 84 L
F 86 N
A 130 S
E 234 D
F 86 N
I 128 Y
A 130 S

superposition of the two models. Together, these sequence differences in the S_2 site are predicted to have significant impact on the binding and kinetics of the substrate-protease interaction.

The S_3 site in these proteases is composed of less than half as many residues as the S_2 pocket (four versus nine). In this context, the three sequence differences in S_3 site change an even greater percentage of the binding site's surface. It should be noted that in the current specificity site designation, the S_3 and S_4 sites share one residue that differs between falcipain-2 and falcipain-1 (N86F) (see Figure 22). Overall there is less hydrophilic functionality lining the site in falcipain-2 (Y78F and N86F), though falcipain-1 has an additional basic functionality (L84H) (see Figure 22). These S_3 site sequence differences are predicted to shrink the substrate binding volume and give rise to a preference for hydrophobic residues for falcipain-1. Falcipain-2 is predicted to have an additional strand at the edge of the beta sheet structure forming part of the S_4 binding site (see Figure 20). Although this region is more likely to play a part in extended specificity, it appears to be the largest global structural difference between falcipain-2 and falcipain-1. This has direct implications for substrate binding. Four out of six

residues in the S_4 site differ between falcipain-2 and falcipain-1. Three of these changes result in gain of hydrophobic functionality in the falcipain-1 S_4 site (see Figure 22). Interestingly, in spite of the sequence differences in the S_4 site, the molecular dimensions of the S_4 pocket differ only by one non-hydrogen atom. It appears that these two proteases do not share substrate specificity at the P4 position, although amino acids with similar volumes may be preferred.

Genome and protein family based drug discovery: falcipain-2 and human homologs in the papain superfamily

A reality of the post-genomic era is access to a seemingly endless array of genome sequences. It is now possible to annotate proteins and analyze the homologies and variations between the pathogen proteins and the human homologs. For oncologic disease, the pathogenic protein(s) may be mutated, up regulated or in the case of tumor supressors, down regulated. While it is possible to imagine small molecule inhibitors of mutated or up regulated proteins (e.g. Gleevec for BCR-ABL (Druker, Talpaz et al. 2001)), down regulated systems are less likely to be amenable to small molecule approaches. In all of these

cases, the impact of a small molecule inhibitor on the related protein targets must be considered. In the case of Gleevec, c-kit and gastrointestinal stromal tumors, an unexpected benefit is found (Strickland, Letson et al. 2001; Tuveson, Willis et al. 2001). However, it is more likely that untoward side effects will result. As is common in the case of infectious disease drug targets, drug resistance of oncoprotein targets can occur by amino acid substitution of residues involved in the drug interaction (Gorre, Mohammed et al. 2001).

Subsite specificity analysis of the falcipain-2 model suggested a number of favorable features for drug design. The dominant feature of the falcipain-2 S_2 site is a deep hydrophobic binding pocket. As judged by peptide substrate binding data, falcipain-2 (Shenai, Sijwali et al. 2000) and the modeling template cathepsin K (Bossard, Tomaszek et al. 1996) share a marked preference P_2 for leucine. The S_3 site is quite small and largely solvent accessible, in keeping with the trend of the papain superfamily. Even the extended specificity S_4 site of falcipain-2 has a potential binding pocket. However, the extended non-prime specificity sites are problematic for drug design because they appear poorly defined structurally and substrate analog binding data shows no preferences in this region (Turk, Guncar et al.

1998). To direct computational small molecule selection calculations and to further understand structure-specificity relationships, we proceeded to analyze the commonalities and distinctions between falcipain-2 and the other plasmodial and human papain-like cysteine proteases.

A multiple sequence alignment based on the available sequences and structures of human cysteine proteases was created and used to assign falcipain-2 residues to substrate specificity sites (see Methods). More than half of the residues on the prime side of the specificity sites are conserved. The few differences have little impact upon site volume, hydrogen bonding or charge. The S_1/S_1' catalytic site including the catalytic dyad, a glycine residue and a number of backbone atoms, is absolutely conserved within this family. This leaves fewer than half of the specificity sites as possible unique structural sites for differential drug design.

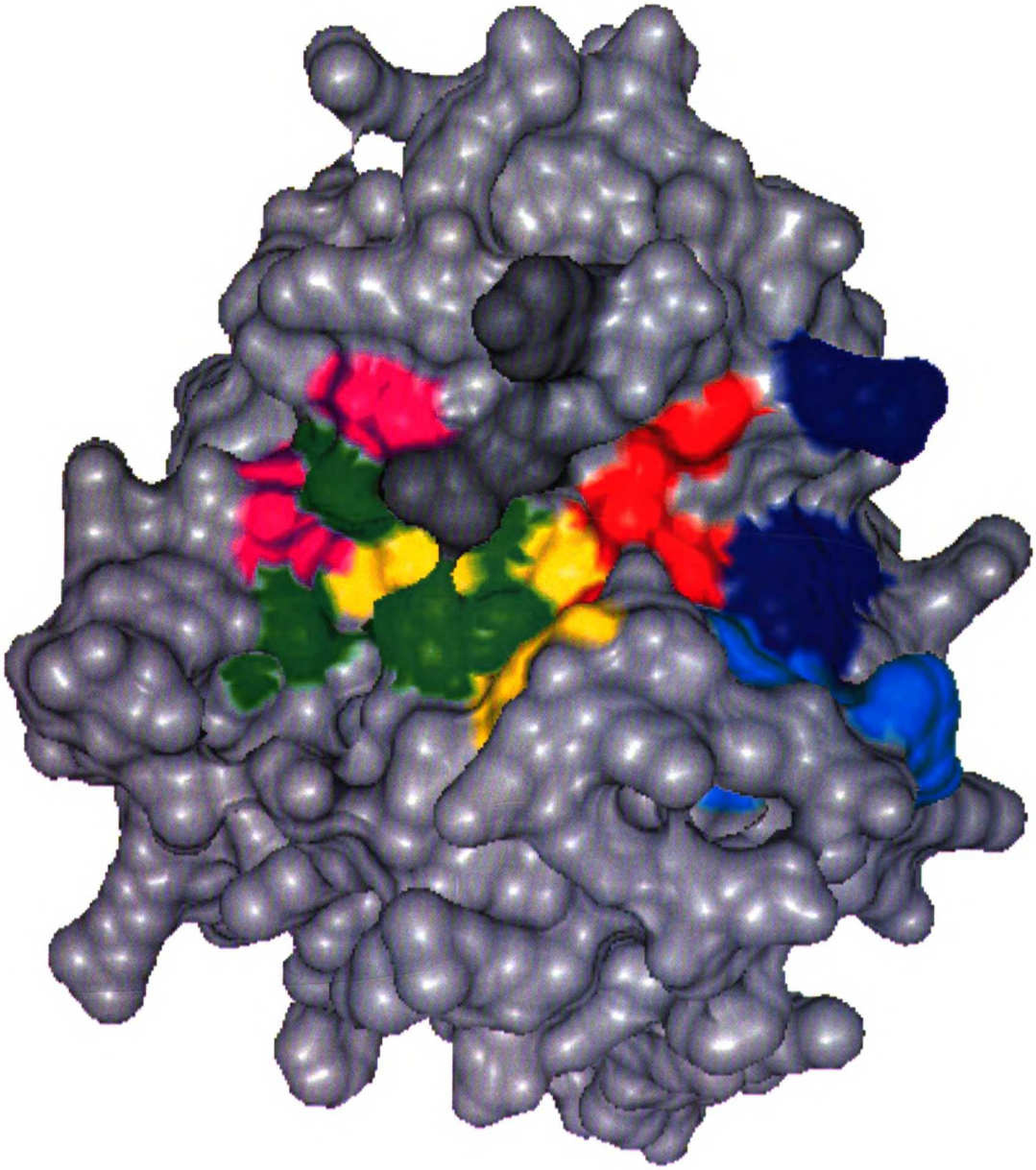
Excluding glycines and main chain atoms, the S_2, S_3 and S_4 sites are variable across the papain cysteine protease family. These sites have diverged during evolution to optimize different functional substrate specificities. Certain sequences exhibit compensating changes, but for nearly all the specificity site sequence positions there exist variations in volume, hydrogen bonding potential and

even charge. Due to the contribution of these specificity sites to substrate binding, such patterns of sequence variation represent the unique functional signature of the papain family. Within the functional variations of a protein family resides an important aspect of protease differential specificity - how changes in sequence affect the binding site volume. If we assume that the backbone positions remain relatively fixed, then mutations to a smaller residue will result in a larger available volume for binding and vice versa. Such unique differences can be exploited with distinct substituents attached to a common small molecule scaffold. In contrast, conserved signatures like the S_1/S_1' catalytic site are problematic for targeted drug design because of their ubiquitous presence within the protein family.

Using the available sequence data we performed a variation of the Evolutionary Trace method (Lichtarge, Bourne et al. 1996), with a JAVA™ implementation of the ET analysis (see Chapter III) (see Methods). Combined with a specificity site annotation by analogy to characterized homologs, the analysis identified amino acids that were unique in falcipain-2 relative to the known human homologs (see Figure 23). A surface-exposed cluster of residues was identified at the boundaries of the S_2, S_3 and S_4 specificity

Figure 23

Unique site analysis of falcipain-2 in context of human homologs. Cathepsins B, C, H, K, L, L2, O, S, Z and stefin B were the human homologs used in this evolutionary analysis. In green are residues unique in falcipain-2 relative to the human sequences. Other colors correspond to the defined specificity sites as represented in Figure 21. See Methods for definition of the specificity sites and the unique site analysis. This figure was generated with a sequence to structure alignment JAVA™ application (Chapter III), CHIMERA (Huang, Couch et al. 1996) and MSMS (Sanner, Olson et al. 1995).



sites. The S_2 site, the main determinant of substrate binding, contained the majority of these unique site positions (three out of six).

The unique site identified in the falcipain-2 specificity sites is solvent accessible, spans two well defined binding pockets and exhibits marked sequence differences relative to human papain family protease homologs. Together this evidence suggested that the identified cluster of residues was a promising candidate drug target site for an antimalarial with minimized specificity towards human homologs of the target. The results of this analysis were directly applied to both the in silico and visual screening steps (see Methods).

Drug discovery results against falcipain-2 compared to falcipain-1.

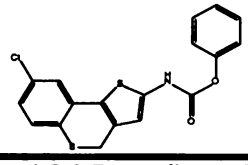
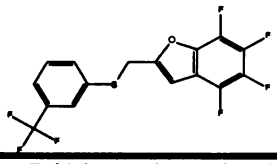
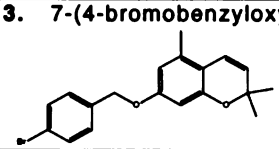
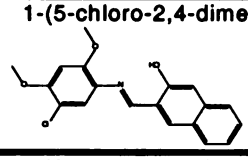
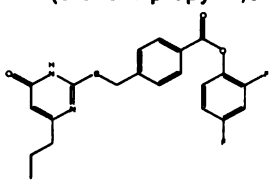
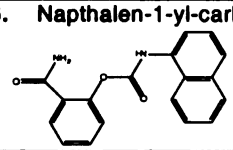
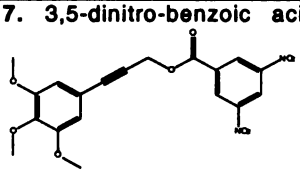
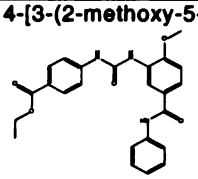
The original falcipain modeling and drug design effort (Ring, Sun et al. 1993) led to three inhibitors with an IC_{50} less than 100 μM. The best compound was a naphthyl-hydrazide, which inhibited the plasmodial protease extract with an IC_{50} of 6 μM in an *in vitro* enzyme assay, and had activity against the parasite in culture at a similar concentration as judged by inhibition of hypoxanthine

uptake. Overall 31 compounds were tested in the original DOCK screen (Ring, Sun et al. 1993), resulting in a 3% hit rate at the < 10 μM cutoff.

The current modeling and drug design effort has had a considerably higher hit rate in terms of active compounds found with the aid of a computational screen. Of the 44 compounds tested, eight had an IC_{50} below 10 μM in an *in vitro* enzyme assay with values ranging from 1 to 7 μM (see Table 3). Three of the eight best compounds against falcipain-2 (**2**, **4** and **7**), were also effective in killing parasites with an IC_{50} of about 20 μM (see Table 3). For these compounds there was complete inhibition of parasite multiplication at 50 μM.

Table 3

Inhibition of falcipain-2 and cultured malaria parasites by compounds identified with a computational screen.

	Enzyme IC_{50}* ** (μM)	Cell Culture IC_{50} ** (μM)
1. (8-chloro-4*H* -1,5-dithia-cyclopenta[a]napthalen-2-yl)-carbamic acid phenyl ester 	1.1 +/- 0.5	94
2. 4,5,6,7-tetrafluoro-2-(3-trifluoromethyl-phenylsulfanylmethyl)-benzofuran 	1.4 +/- 0.6	20
3. 7-(4-bromobenzyloxy)-2,2,5-trimethyl-2*H* -chromene 	2.5 +/- 1.3	no inhibition
4. 1-(5-chloro-2,4-dimethoxyphenyliminomethyl)-2-naphthol 	3.5 +/- 0.6	25
5. 4-(6-oxo-4-propyl-1,6-dihydroprymidin-2-ylsulfanylmethyl)-benzoic acid 2,4-difluoro-phenyl ester 	4.1 +/- 1.3	ND
6. Napthalen-1-yl-carbamic acid 2-carbamoyl-phenyl ester 	4.7 +/- 2.8	103
7. 3,5-dinitro-benzoic acid 3-(3,4,5-trimethoxy-phenyl)-prop-2-ynyl-ester 	6.4 +/- 1.1	21
8. 4-[3-(2-methoxy-5-phenylcarbamoyl-phenyl)-ureido]-benzoic acid ethyl ester 	6.9 +/- 2.6	ND

* Leupeptin, the enzyme assay positive control, had an IC_{50} of 50 nM.
** See Methods for details.

Discussion

There are several advantages to pursuing a family of proteins as drug discovery targets, including: 1) compounds related to inhibitors of one family member are likely to be active against other members; 2) the structure of one member provides substantial insights into the structure and function of other members; 3) assay development can proceed in parallel; and 4) experience developed for one target is frequently relevant to the homologous targets. Confounding these advantages, inhibitor specificity can be a significant challenge and the relatedness of the targets means that the inhibitors, especially those that are mechanistic in nature, are likely to have several distinct activities.

In the case of cysteine proteases, not only are there thousands of cysteine protease sequences in the sequence databases, but this ubiquitous sequence family also has tens of well-determined crystal structures complexed with inhibitors. This is a favorable situation for modeling, predicting substrate specificity and forming inhibitor structure-activity relationships.

Based on our evolutionary analysis, the S_2, S_3 and S_4 specificity sites of falcipain-2 exhibit unique functional differences that can be targeted with drug design. In contrast, conserved signatures traditionally targeted with mechanistic inhibitors, like the S_1/S_1' catalytic site, logically lead to the selection of compounds with specificity towards human homologs of the drug target. In addition, the non-prime sites encompassing the unique falcipain-2 site should lead to decreased drug side effects. We predict however, that for instances involving a target whose interactions are strictly selected for and where the binding partners (small or macromolecules) also exhibit chemical conservation, the present analysis will not guarantee a unique cluster of residues. Nevertheless, for many protein families, functional speciation can be observed in the coevolution of binding interfaces and how side-chain variation at the interfaces occurs in a correlated manner (Goh, Bogan et al. 2000). All instances of speciation of function result in differences that can be exploited to direct drug design efforts.

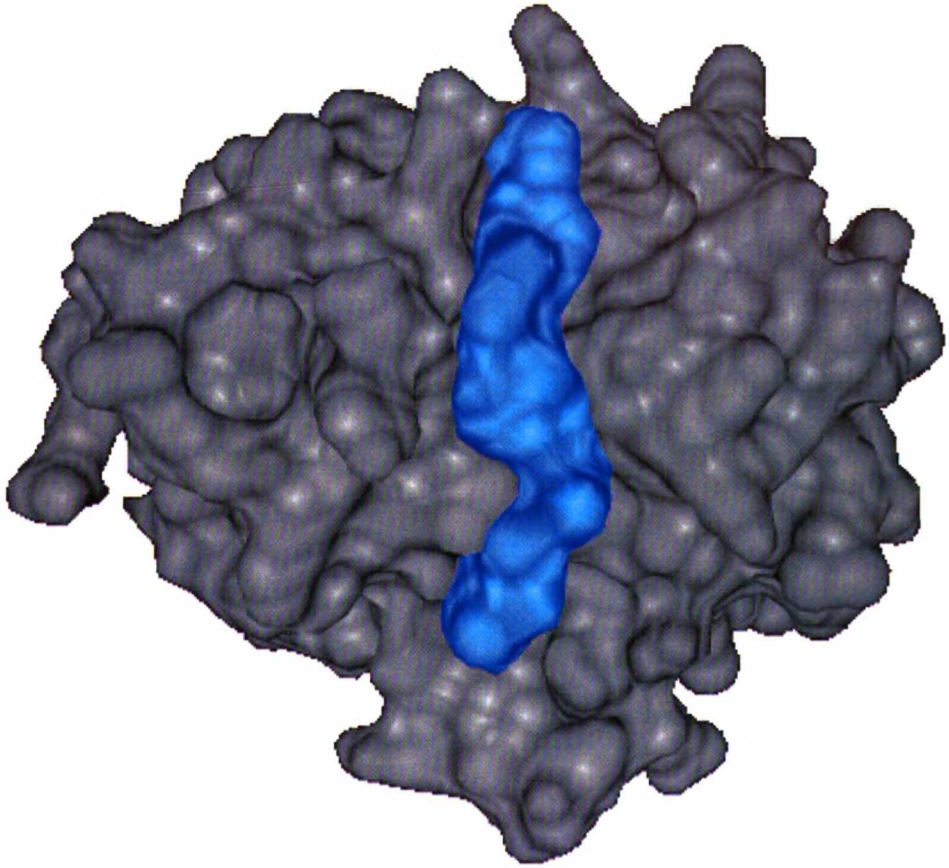
Comparison of the structural sites targeted with computational screening shows important differences between falcipain-1 and falcipain-2. For falcipain-1 the most active compound was predicted to bind to the $S_2/S_1/S_1'$ sites

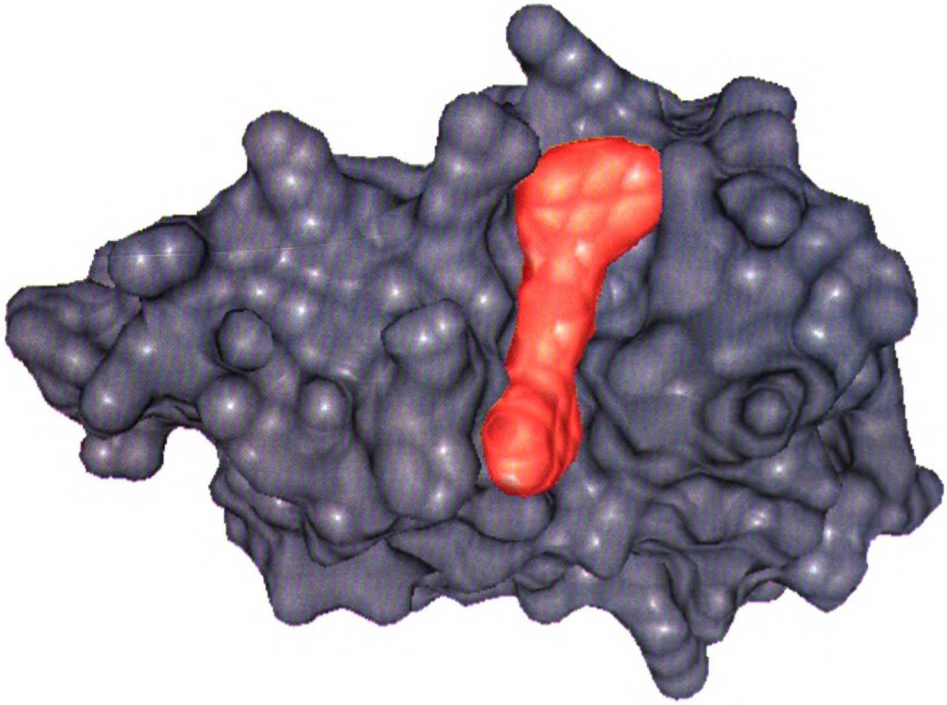
(see Figure 24). Given that S_1 and S_1' are the conserved signature of the papain family, this inhibitor has the potential to cross-react with human cysteine proteases. The set of inhibitors generated by virtually screening the falcipain-2 model produced eight diverse compounds, all selected to bind the unique functional signature of the S_2 and extended non-prime sites. The common feature of many of these compounds, including the falcipain-1 inhibitor, is a pseudo-peptide backbone of length 2-4 atoms adopting a planar conformation owing to the double bonds and the two flanking functionalized ring systems. We continue to believe that the length of the linker and chemical substituents on the functional groups are what determine the specificity and uniqueness of the target-inhibitor interaction (Li, Chen et al. 1994; Li, Chen et al. 1996).

It is difficult to extrapolate the present cell culture results to the antimalarial action of other known falcipain inhibitors because earlier cellular assays varied in method. In addition, the actual concentration of compound reaching the parasite food vacuole in the cell culture experiments is likely to be affected by permeability through the multiple cell membranes involved. All of the compounds tested in this study had molecular weights less than 350 Da, and all were soluble in water at

Figure 24

Inhibitor binding modes of falcipain-1 compared to falcipain-2. Models show the predicted binding modes of the best inhibitors for falcipain-1 (orange, left) and falcipain-2 (blue, right). The falcipain-1 model structure is shown with the predicted binding mode of its best symmetric acyl-hydrazide inhibitor from the work of Ring et al (Ring, Sun et al. 1993). The falcipain-2 model is shown with the predicted binding modes for the top eight active compounds. The figure was generated with MSMS (Sanner, Olson et al. 1995; Sanner, Olson et al. 1996) and the MSV software (Sanner, Olson et al. 1996).





appreciable concentrations. Nevertheless, it cannot be ruled out that the cell culture IC_{50} measurements in our experiments do not reflect the effective concentration of compound at the target site, presumably the parasite food vacuole.

There are some factors convoluting the results of both the *in vitro* enzyme and the *in vivo* culture inhibition assays. An important difference in the drug discovery effort against falcipain-1 compared to falcipain-2 is the size of the chemical database used in the computational screen. The Fine Chemicals Directory used in 1993 contained 55313 compounds, whereas the version of the Available Chemical Directory used in the current drug discovery effort consisted of 195419 commercially available compounds. This 4-fold increase in database size results in significantly greater chemical diversity available for computational screening and drug discovery. The percentage sequence identity to the modeling target, a standard measure of model accuracy, is predicted to have had a negligible effect in this case. The percentage sequence identity between falcipain-1 and papain and actinidin was 33 %, while that of falcipain-2 to cathepsin K was 35 % (see Methods). Substantially more sequence similarity was

observed in the region around the active site that should be most important to drug design efforts.

The most powerful convolution in the falcipain-1 drug discovery effort was the assay of enzyme inhibition performed using a parasite extract now known to be primarily composed of falcipain-2. Knowledge about the expected and associated phenotypes has considerably increased in the past few years - assessing the *in vivo* inhibition phenotype began with a general metabolism assay, continued with food vacuole swelling as exhibited by broad-spectrum inhibitors, and now with multiple homologous targets has returned to more general assays of parasite health and development. Significantly, technological improvements in modeling, structure analysis and docking, have become combined with the accumulation of sequences, crystal structures, and available small molecules. An important remaining rate limiting step in post-genomic drug discovery is knowledge about the target. Such knowledge includes the cellular and disease contexts, other gene products modulating the targets' function, as well as the functional family it belongs to within a specific genome and beyond.

Presently the function of falcipain-1 remains unknown and the current analysis may serve as a lead in the search

for its endogenous targets. The sites responsible for substrate specificity in falcipain-1 and falcipain-2 are notably different given the high degree of sequence similarity. It is postulated that these two cysteine proteases have different endogenous target sequences and therefore different *in vivo* functions.

Our drug discovery effort against falcipain-2 has resulted in eight compounds with activities < 7 μM, three of which kill parasites in a cell culture assay. Presumably, their potency can be increased through a combined medicinal and computational chemistry effort directed at the unique site of falcipan-2. Most importantly, nearly all antimalarials in current use have pronounced side effects and/or have encountered plasmodial drug resistance. The molecules identified in the falcipain-2 drug discovery effort represent diverse chemical scaffolds and functionalities relative to the known antimalarial drugs (see Table 4). It follows, that the structures we have identified as falcipain-2 inhibitors provide new avenues for antimalarial drug development with potential for minimized toxicity and drug resistance.

Table 4

Chemical structures and properties of prescribed antimalarial drugs.

Name	Resistance [32]	Toxicity [32]	Compound
quinine	Yes	delayed 'glaucoma'	
chloroquine	Yes	heart liver	
hydroxychloroquine	Yes	retina macula	
primaquine	Yes	anemia	
amodiaquine	like chloroquine	agranulocytosis with long term treatment	
mefloquine	Yes	cardiotoxicitiy vivid dreams psychosis	
proguanil	Yes DHFR	ulcers, alopecia	
atovaquone (with proguanil)	Yes selective cytochrome b	minor	
pyrimethamine (with sulfadoxine or sulfonylbisbenzenamine)	Yes DHFR	severe skin disease	
sulfadoxine (with pyrimethamine)	Yes	severe skin disease	
sulfonylbisbenzenamine (with pyrimethamine)		anemia, allergy,fever	
doxycycline (with quinine)		sun sensitivity	
Artemisinin		CNS toxicity	
halofantrine		cardiotoxicitiy	
Pyronaridine		teratogen	

Conclusions

Based on modeling and drug design we have explored the effects of the association between a gene product and its phenotype in the context of a gene family. Discovery of the new sequences and their experimental confirmation as targets immediately led to new models of a new target. Higher expectations for success were set for drug design efforts, specifically structure-activity correlations and improved specificity for the parasite enzyme, both in culture and in animal tests. Based on the current drug design effort and other examples of protease inhibition, a primary cause of side effects is accumulation of undesirable substrates. In molecular terms such side effects are a signal for the presence of genes potentially unrelated to the target phenotype (e.g. ACE inhibition leads to accumulation of bradykinin and substance P (Emanueli, Grady et al. 1998)). In the case of malaria infection in humans, the side effects of broad papain family inhibitors would most likely mean accumulation of many substrates, the majority of which are host proteins.

Biological phenotypes are fundamentally complex due to the presence of back-up functionalities, possibly

distributed across cell types, tissues and time, as well as clearance and other protective mechanisms. These pitfalls of drug development illustrate the detailed knowledge required for pursuing targets even with established phenotypes. A gene family perspective can lead to unique structural sites relative to the 'host' or 'pathogen' families. Knowledge of gene families can aid in the identification of the inhibitory spectrum of a molecule and provide the insight and unanticipated auxiliary functions of the original target or back-up functions subsumed by family members.

Materials and Methods

Homology methods and structural modeling

A structural model of falcipain-2 was constructed by homology to other members of the papain cysteine protease subfamily. In order to identify a structural template, the protein structure database (PDB) was searched for falcipain-2 homologs using the PSIBLAST algorithm (Altschul, Madden et al. 1997). The closest homologs were the cathepsin K zymogen (e-value of 3E-43 over 321 residues), the cathepsin L zymogen (2E-41 over 326 residues), the caricain zymogen (4E-41 over 343 residues), and the ginger rhizome cysteine protease (1E-38 over 220 residues). Typically, the best template for modeling corresponds to the sequence with the longest significant alignment and the highest score in the mature protease region. At a per-residue level, human cathepsin K was found to be 39 % identical in the mature region of the protease (35 % identity over all aligned residues).

A model of falcipain-2 based on the cathepsin K zymogen structure (PDB code: 1BY8, resolution 2.6 Å) was built using MODELLER (Sanchez and Sali 1997). This software

derives distance and angle constraints based on conserved sequence features in the alignment and structural features of the template. Given a correct alignment, sequences that share 40 % identity are expected to align within 1 Å RMS over 90 % of their residues, approximately the accuracy expected in the present model. Falcipain-2 has an additional predicted disulfide bond relative to cathepsin K, and this disulfide bridge was added with the modeling software SYBYL (TRIPOS corp.). The model structure was refined at the side-chain level using the backbone-dependent side-chain rotamer library algorithm SCWRL (Bower, Cohen et al. 1997). For pairs of sequences with gaps inserted to yield an alignment with identical residues in 30-40% of the positions and using a template structure determined using 2 Å resolution x-ray data, SCWRL predicts the χ_1 side-chain angles with an accuracy of 65%. All identical aligned residues were fixed in their template conformation. Of the eight residues with unlikely conformations identified by SCWRL, all were distant from the active site and occurred in regions of insertions relative to the structural template.

Annotation of Specificity Subsites and Unique Site Analysis

The substrate specificity sites in the falcipain-2 model structure were identified by analogy to the extensive family of papain-like cysteine proteases. Following the nomenclature of Berger and Schechter (Berger and Schechter 1970) and sequence alignments to known papain family crystal structures, the falcipain-2 S_4 to S_2' substrate side-chain binding sites on either side of the scissile amide bond were identified. A more extensive multiple sequence alignment was built using CLUSTALW (Thompson, Higgins et al. 1994) and edited to include the alignments derived by structure alone.

As we seek inhibitors that are unlikely to be active against human proteases from the papain family, we performed a variation of the Evolutionary Trace method (Lichtarge, Bourne et al. 1996) on falcipain-2 and its human homologs. The variation consists of restricting the sequence data to a subset of the full sequences, corresponding to the aligned specificity sites. The definition of residue conservation and subfamily comparison was modified, by considering only amino acids that were unique in human sequences relative to the falcipain-2

target. Finally, residue similarity filters (BLOSUM62 (Henikoff and Henikoff 1993)) and coloring by groups were applied, to analyze the unique positions in terms of volume, hydrogen bonding and charge properties. All of the above functions and the resulting mapping of phylogenetic and sequence data onto the falcipain-2 model structure (Figure 23) were performed with a JAVA™ application (Chapter III).

Docking

DOCK 4.0 (Ewing, Makino et al. 2001) was used to screen the falcipain-2 sites against the Available Chemicals Directory release 97.2 containing 195419 unique compounds (MDL Inc). The screening procedure took about six weeks of CPU time on a 4 processor MIPS R12000 SGI server. 5000 compounds were saved from the energy-scoring scheme, and 5000 from the shape-scoring scheme. Visual selection of these hits was performed in duplicate, to arrive at a set of 160 compounds in an unbiased fashion. This selection step relied on knowledge of the specificity sites unique in falcipain-2 relative to known human sequences, as well as standard drug-like properties of small molecules including: hydrophobicity, molecular weight, and absence of chemical

functionalities with tendencies to form covalent adducts with amino acid side-chains.

Falcipain-2 enzyme assays

A set of 44 compounds manually selected from the DOCK computational screen of the falcipain-2 active site was tested in a fluorescence-based assay against recombinant falcipain-2. Recombinant falcipain-2 was prepared (Shenai, Sijwali et al. 2000) and the falcipain-2 fluorescence-based assay performed as previously described (Rosenthal, Wollish et al. 1991). All compounds were dissolved in DMSO to make a 10 mM stock solution. Each compound was incubated with the enzyme in 0.1 M sodium acetate (pH 5.5) and 10 mM dithiothreitol (DTT) for 30 minutes at room temperature before addition of the substrate benzyloxycarbonyl-Phe-Arg-7-amino-4-methyl-coumarin (Z-Phe-Arg-AMC). The fluorescence caused by the cleavage of the substrate was monitored continuously over 30 minutes with a Fluoroskan II spectrofluorometer (Labsystems). The rate of hydrolysis of Z-Phe-Arg-AMC in the presence of the compounds was compared with the rates of hydrolysis in the negative (equivalent volume of dimethyl sulfoxide (DMSO)) and positive (100 μM leupeptin) controls. Using the PRISM 3.0 software (Graphpad

Software Inc.), the 50% inhibitory concentration of each compound (IC_{50}) was determined from plots of falcipain-2 activity inhibition over a series of compound concentrations. Initial fluorescent assay screens were carried out for 44 DOCK compounds and those with IC_{50}'s below 10 μM were selected for further testing.

Cell Culture Assays

The six best compounds, **1-4**, **6** and **7** (see Table 3) were selected for characterization in a cell-based assay. Final concentrations of compounds in the parasite cultures were 100 μM, 50 μM, 25 μM and 10 μM and the final concentration of the DMSO control was 1%. W2 strain *P. falciparum* parasites were cultured with human erythrocytes at 2% hematocrit in RPMI-1640 medium supplemented with 10% heat inactivated human serum (Rosenthal, Wollish et al. 1991). Parasite synchrony was maintained by serial treatments with 5% sorbitol (Lambros and Vanderberg 1979). In order to assess the effects of inhibitors on parasite development, *P. falciparum* parasites were incubated for 48 hours with different concentrations of compound added from 100X stocks in DMSO (Rosenthal, Wollish et al. 1991). The experiment was started at the synchronized young ring stage

and continued until the control cultures contained nearly all new ring stage parasites (48 hours). Giemsa-stained smears were made at 24 and 48 hours. At 24 hours parasite morphology was evaluated and at 48 hours the number of new ring forms per 1000 erythrocytes were counted and compared to control cultures incubated with DMSO. IC_{50}'s for compounds **1-4**, **6** and **7** were calculated using the PRISM 3.0 software.

Acknowledgements

We are grateful to Xiaohui Du, Florence Horn, Barney May and Elaine Meng for their comments and advice. This work was supported by grants from the Burroughs-Wellcome Fund and the National Institute of Health.

Chapter III

JEvTrace: refinement and variations of the Evolutionary Trace

Abstract

Motivation

From an analysis of gene families within and across genomes, it is clear that speciation is a useful concept for understanding structural and functional variation at the molecular and organismal level. However, some of the functional speciation within these gene families is lost when comparisons are limited to standard multiple sequence alignments.

The Evolutionary Trace (ET) (Lichtarge, Bourne et al. 1996) was developed to combine multiple sequence, phylogenetic and structural data to identify functional sites in proteins and to extract detailed insights into the evolution of functional surfaces of macromolecules. The ET method has been successfully applied in a number of biological systems including signaling proteins and receptors (Lichtarge, Bourne et al. 1996; Lichtarge, Yamamoto et al. 1997; Landgraf, Fischer et al. 1999; Pritchard and Dufton 1999; Gouldson, Higgs et al. 2000; Innis, Shi et al. 2000; Sowa, He et al. 2000; Aloy, Querol et al. 2001; Joachimiak, Chang et al. 2001; Sowa, He et al.

2001; Lichtarge, Sowa et al. 2002). However, no public and user friendly implementation of this method has been available thus far. Due to the complexity of the input data and the informatics problems of function and specificity discovery, there exist useful variations of the ET method. In addition, the complexity of the output of ET data and the desire to couple interpretation of different data types heightens the need for a flexible and graphical user interface. We have addressed this problem by developing data structures that represent graphical results of multiple sequence alignment analyses with a phylogenetic and structural perspective.

Results

We have implemented the ET in a JAVA™ application, JEvTrace. The implementation allows users to access the underlying data and results of ET analysis. A number of variations of the ET can be performed on any combinations of nodes of the phylogenetic tree. Since protein families and phylogeny represent complex data with statistical outliers and special cases, this flexible approach to the ET allows more extensive and detailed mining of evolutionary sequence relationships, and remedies

limitations present in the original implementation. Function discovery with JEvTrace is demonstrated on the example of two proteins with recently determined crystal structures: the protein YlxR from *Streptococcus pneumoniae* with a predicted RNA-binding function (Osipiuk, Gornicki et al. 2001), and a *Haeomophilus influenzae* protein with unknown function, YbaK (Zhang, Huang et al. 2000).

To facilitate analysis and storage of results we propose a multiple sequence alignment (MSA) coloring format that relies on the inherent structure of MSA's, including the evolutionary features of protein sequence families. The Sequence Coloring Format (SCF version 1.0) is optimized for storage and accessibility criteria, and enables flat file storage of any and all possible colored selections of a MSA.

Introduction

Whole genome analyses have enabled the study of gene families within and between species. From computational and experimental studies of these genomes and gene families, new perspectives are emerging on evolution of specificity and cellular metabolic organization. However, these efforts remain limited by our ability to accurately annotate gene function. In yeast, the number of ORFs with assigned functions by sequence similarity based methods was 43% (Mewes, Albermann et al. 1997). With the inclusion of extensive experimental data this value is approaching 70% (Mewes, Frishman et al. 2002). Meanwhile, a search of the PDB for the keyword 'unknown function' retrieved 31 protein structures. Many of these structures are the results of structural genomics initiatives. As this number is likely to grow, it has become more important to develop computational tools to intuit function from analysis of sequence information in the context of structure.

Assigning function by sequence homology alone presents a number of caveats, including the occurrence of structurally homologous enzymes catalyzing different reactions (Gerlt and Babbitt 2000) and the propagation of

error through successive rounds of sequence annotation (Brenner 1999). Conversely, assigning function by structure alone can also be daunting, even if one ignores the implicit selection bias in structure databases relative to sequence databases. Analysis of the CATH database revealed that while function was conserved in nearly 51% of enzyme families, function had diverged considerably in highly populated families (Pearl, Todd et al. 2000). This consequence has direct implications for structure based function predictions using threading algorithms (Jones, Tress et al. 1999; Panchenko, Marchler-Bauer et al. 1999). Another serious complication in the structure based function discovery problem is the intrinsic limit on our ability to compare distantly related sequences and to recognize the role of specific residue subsets in polyfunctional proteins. It can be difficult to recognize if a distantly related homolog belongs to a superfamily with a functional supersite (Russell, Sasieni et al. 1998) or whether that particular structural scaffold accommodates multiple functional sites, as with the G-proteins (Lichtarge, Bourne et al. 1996).

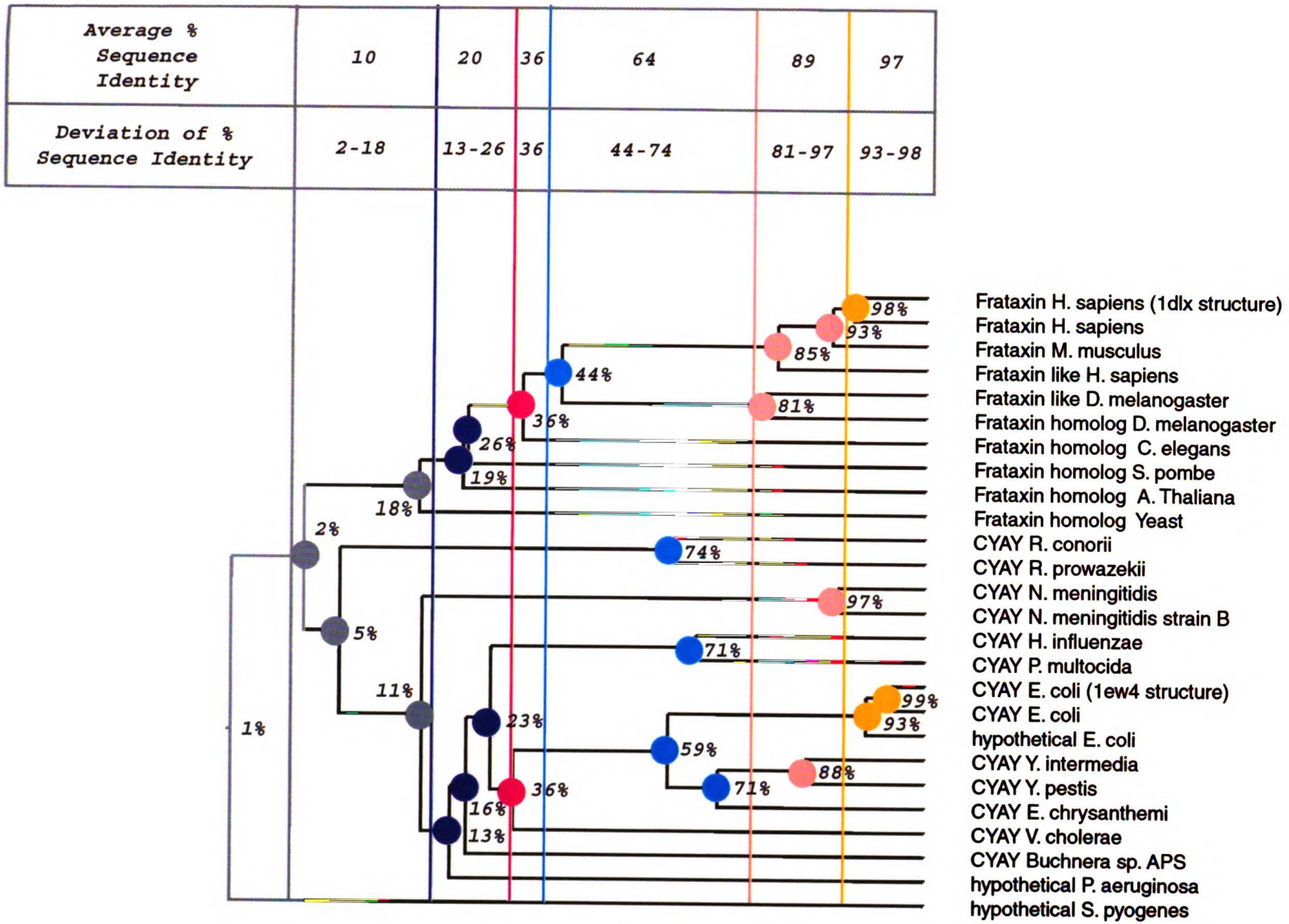
It follows that similarity free function prediction methods are especially desirable. Marcotte et al employed correlated evolution, correlated mRNA profiles and patterns

of domain fusion for genome-wide function prediction (Marcotte, Pellegrini et al. 1999). A method based on local gene order of orthologous genes has been proposed (Kolesov, Mewes et al. 2001). Protein-protein interactions have been used to assign function with surprising success (Hishigaki, Nakai et al. 2001) and functional descriptors have been used to search structure space (Di Gennaro, Siew et al. 2001). However, the predictive capabilities of these methods when used individually remain unsatisfactory.

ET presumes that the branchpoints separating subclades of a phylogenetic tree can specify molecular speciation events, and hence evolutionary selection of amino acids. Thus, nodes can mark points in evolution where a protein gains, modifies or loses a binding or catalytic function (Lichtarge, Bourne et al. 1996). The original ET method relies on a partitioning of the phylogeny. This procedure results in sets of nodes at different levels of Percent (sequence) Identity Cutoffs (PIC) (Du and Alkorta 1994). However, since phylogenies often contain extreme branches due to distant homologs or rapid speciation, pairs of protein family members are not represented uniformly across the sequence identity range. This is reflected in a skewed topology of the phylogeny, e.g. the *P. aeruginosa* and *S. pyogenes* hypothetical proteins at the bottom of Figure 25.

Figure 25

A phylogenetic tree of the frataxin family (Cho, Lee et al. 2000) detailing the presence of distant sequence homologs within a phylogeny. The MSA and dendogram of the family were constructed as described in Methods. Partitions of the phylogeny are shown as colored vertical bars. Each partition of the phylogeny corresponds to an interval of percent sequence identity. The percent sequence identity for the selected subclades is shown on the phylogeny.


Average % Sequence Identity
10
20
36
64
89
97
Deviation of % Sequence Identity
2-18
13-26
36
44-74
81-97
93-98
98%
93%
85%
44%
81%
36%
26%
19%
18%
2%
74%
97%
5%
71%
11%
99%
93%
1%
23%
59%
88%
36%
71%
16%
13%
Frataxin H. sapiens (1dlx structure)
Frataxin H. sapiens
Frataxin M. musculus
Frataxin like H. sapiens
Frataxin like D. melanogaster
Frataxin homolog D. melanogaster
Frataxin homolog C. elegans
Frataxin homolog S. pombe
Frataxin homolog A. Thaliana
Frataxin homolog Yeast
CYAY R. conorii
CYAY R. prowazekii
CYAY N. meningitidis
CYAY N. meningitidis strain B
CYAY H. influenzae
CYAY P. multocida
CYAY E. coli (1ew4 structure)
CYAY E. coli
hypothetical E. coli
CYAY Y. intermedia
CYAY Y. pestis
CYAY E. chrysanthemi
CYAY V. cholerae
CYAY Buchnera sp. APS
hypothetical P. aeruginosa
hypothetical S. pyogenes

Hence, the PIC cutoffs correspond to intervals of percent sequence identity, and the greater the number and/or magnitude of outliers in the family, the larger the percent identity interval (see Figure 25). Presence of these outliers effects multiple alignment and phylogenetic models, and in ET analysis can misrepresent the functional variability at the presumed PIC level. This issue has been addressed by normalizing the score in the ET method with sequence variability and sequence uniqueness measures (Landgraf, Fischer et al. 1999). However, numerical normalization reduces the problem to one of sequence analysis, in effect disregarding evolutionary aspects. In the case of distant subclades, ET analysis of appropriately chosen subclades of the phylogeny will have the desired normalization effect. This approach can be used to correct for positional variability, sequence representation bias and non-uniform phylogenetic topologies.

Another limitation of the original ET method was the definition of invariant and neutral position types. Lichtarge et al (Lichtarge, Bourne et al. 1996) recognized that with growing sequence databases, the strict definitions of invariance as a total lack of invariance, and neutrality as invariance discrepancy in even one family, were destined to evolve. Inherently, the functional

resolution in the ET method relies on an optimization based on ET results from multiple partitions, each corresponding to unique definitions of subclade invariance. This optimization has so far evaded automation, with users having to resort to manipulation of the underlying data and cycles of ET analysis and visual inspection of the results mapped to protein structures.

Aside from manually filtering and pruning the data, there has been no simple way of controlling which subclades of the protein family are used in the analysis. An elegant solution to this problem is to allow the user to access all possible subclades represented by the phylogenetic tree. In this way a number of ET variations can be performed, extending the analysis to multiple views of protein family evolutionary data.

JEvTrace is one possible implementation of protein family analysis. Such analyses, which include experimental techniques such as alanine scanning (Wells 1991) and computational techniques such as MSA coloring schemes (Taylor 1997), attempt to organize the massive amounts of sequence and structure data. The results introduce the problem of choice of strategies to identify biologically meaningful patterns. In general, sequence and structure alignments are frequently used to sort features of gene

family data. Analysis of alignments provides coherence to the understanding of biological data, especially from the perspective of distinct features that may explain the unique functional attributes of an individual entry. As is common in ET analysis, these features may extend over sequentially or spatially clustered sets of nucleotides or amino acids and patterns are frequently difficult to identify without a form of color coding. Obviously, color coding exploits our cognitive pattern recognition skills - skills that have been difficult to replicate algorithmically.

Results

An example of a protein with unknown function is 1G2R, representing the YlxR gene from *Streptococcus pneumoniae* (Osipiuk et al 2001). YlxR belongs to the putative nusA/infB operon in *S. pneumoniae*. The operon contains seven genes, three of which are conserved in other bacteria: RbfA, nusA (with its well characterized gene product IF2 (Grill, Moll et al. 2001)), and infB. All three of these proteins are involved in translation and ribosomal function during cold shock response (Bae, Xia et al. 2000). YlxR has been assigned to COG 2740 (Tatusov, Natale et al. 2001), which contains the conserved amino acid motif GRGA(Y/W). The proteins of COG 2740 are predicted to be nucleotide-binding proteins implicated in transcription termination. Several features of the structure, including the conserved and appropriately spaced arginines that could form a characteristic positively charged surface patch (Figure 26 B,C), and a large bent groove reminiscent of other RNA-binding structures, supported the argument that YlxR is a RNA-binding protein.

Structures of proteins in complexes with small molecules have led to and confirmed predictions of protein

Figure 26

JEvTrace analysis of the YlxR protein family. Absolutely conserved positions are colored red. The two major subclades are labeled S_1 and S_2 and the location of the *S. pneumoniae* YlxR (PDB:1G2R) in the phylogeny is shown. B shows the results of subclade trace analysis of the five minor subclades, highlighted with black squares on the phylogeny in A. In B the per-residue score corresponds to the color coding as in Figure 27. C highlights the subclade sequence conservation comparison between the two major subclades, which are highlighted with large blue and orange circles on the phylogeny in A. In C, the residue coloring corresponds to the blue and orange designation of the subclades. Graphics of the molecular surfaces were created with Chimera (Huang 1996) and MSMS (Sanner, Olson et al. 1996) using the SCF format to import JEvTrace results. Graphics of the phylogeny were created with JEvTrace.

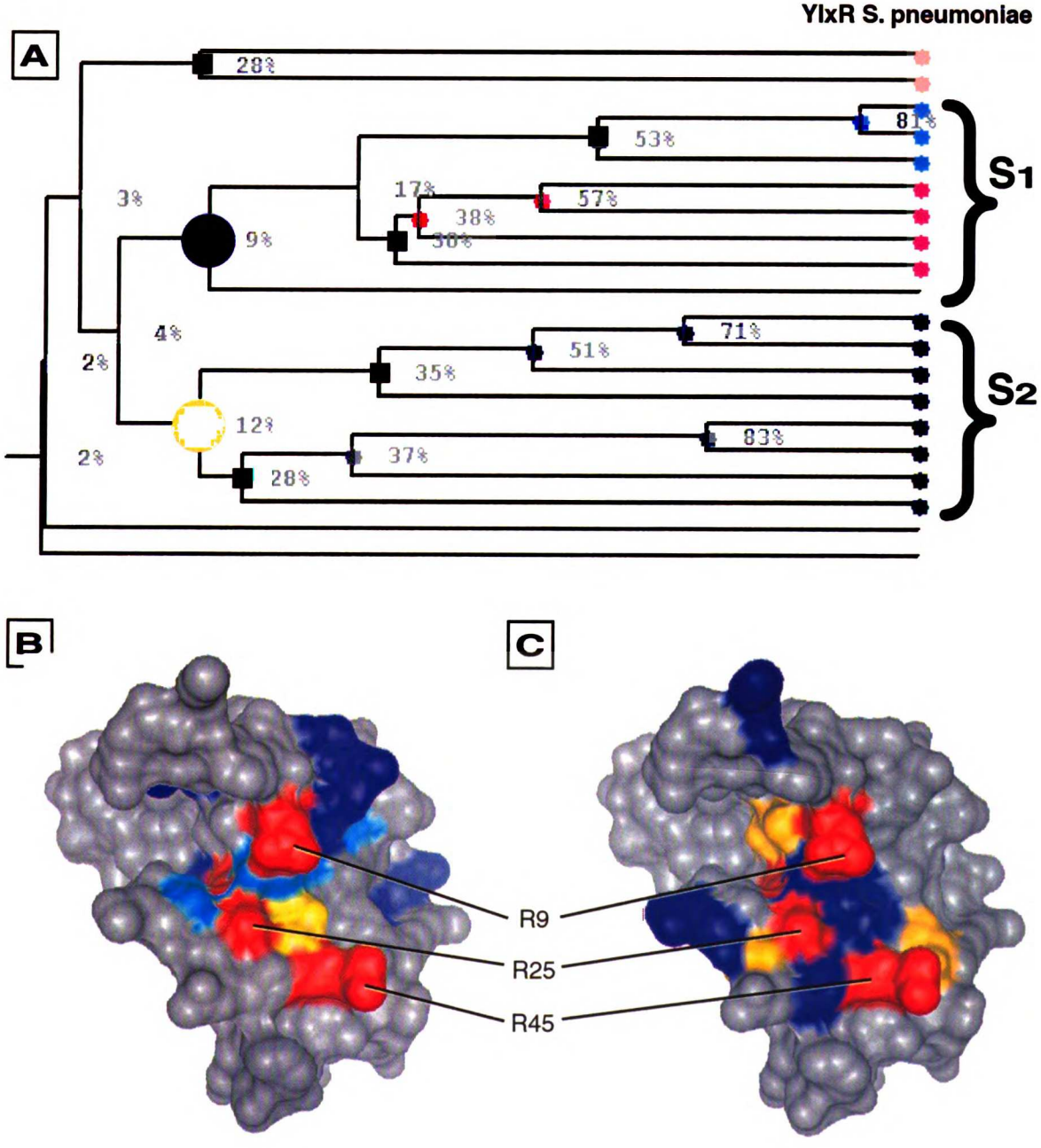

YlxR S. pneumoniae
A
28%
81%
53%
17%
57%
38%
30%
3%
9%
S1
4%
71%
51%
2%
35%
12%
83%
2%
37%
28%
S2
B
C
R9
R25
R45

Figure 27

A description of the scoring scheme and coloring scale used in JEvTrace. The score for a given position is calculated as the pairwise sum of subclade invariance within a given partition or set of nodes. In the event of numerous invariant subclade pairs, the score is normalized to an integer interval corresponding to colors in the color scale.

Minimum Invariance

Maximum Invariance

Pairs of Subfamilies (case < 8)	Pairwise Subfamily Score Mapped to Integer Interval
1 pair of subfamilies	1 pair of subfamilies
2 pairs	
3	
4	
5	
6	
7	Maximum pairs of subfamilies

function. The structure of YlxR is complexed with three sulfate ions, two of which are bound to R25 and R45, and the third to a lysine pair K62 and K63. It has been observed that the distances between the sulfate ions correspond to distances between phosphate groups in a RNA duplex (Osipiuk, Gornicki et al. 2001). The predicted binding site also fully encompasses two out of three of the sulfate ions and borders the third.

The protein family retrieved by PSIBLAST (Altschul, Madden et al. 1997) consists of 20 unique sequences. The hypothetical ancestor sequence has three conserved arginines. Although this is a relatively small family for ET analysis, a simple multiple sequence alignment suggests that only the arginine of the conserved GRGA(Y/W) motif is absolutely conserved. However, since the key arginines associated with the predicted function are absolutely conserved, this implies that the predicted RNA-binding function is conserved across this family.

JEvTrace subclade trace analysis was performed on all of the subclades of this family, ranging from 28 to 53 % sequence identity (Figure 26). After filtering out residues buried from solvent, the following residues were identified in the vicinity of the conserved arginines (square brackets indicate the conserved arginine): K10, V12, V13, S14, K20

[R9], G40 [R25], G46 [R25], 48Y [R25, R45], and K30 and E31 [R45]. At least eight residues form a spatial cluster in the vicinity of the conserved arginines. The residues K10, the backbone of S11, V12, V12, S14, V17, G40, E55, K63, and 64V, from the kinked C-terminal helix, a beta-turn, and parts of the beta-sheet, potentially define a binding epitope (Figure 26B). The epitope includes a collection of hydrophobic interactions that have been correlated with the evolution of residues forming a surface epitope in the vicinity of R9 (Figure 26B).

There appear to be two distinct subclades within the YlxR sequence family, both consisting of proteins with unknown or uncertain function (S_1 and S_2, Figure 26). A subclade comparison was performed to analyze the conserved residues of these two subclades. Such a comparison is useful when a protein family has few representatives or limited evolutionary diversity i.e. few subclades. It appears that independent sets of residues are conserved. All of these residues are in the vicinity of one or more of the conserved arginines, and define a slightly larger and differently oriented surface epitope (Figure 26C) than in the JEvTrace subclade trace analysis (Figure 26B). We propose that the conserved residues modulate the specificity of the predicted RNA interaction, and that the

two subclades correspond to specificity subtypes within the larger family, possibly with unique functional features. The residues not identified by subclade trace analysis but appearing in the subclade comparison, are responsible for a finer level of molecular specificity. In this case of a predicted protein function, JEvTrace analysis presented direct evidence for additional binding functions and highlighted the presence of potential subtypes in the RNA binding specificity.

Another interesting family of unknown function is the bacterial YbaK proteins. A structure of the homolog from *Haemophilus influenzae* has been solved (Zhang, Huang et al. 2000). This gene product has been proposed to serve as a regulatory protein (Burns and Beacham 1986; Bensing and Dunny 1993). Analysis of the sequence family in the context of the structure revealed one conserved residue K46 in a small putative binding site (Zhang, Huang et al. 2000). The YbaK fold is related to a circular permutation and truncation of the C-lectin fold. However, a saccharide binding function for YbaK is unlikely due to a small putative binding site and lack of saccharide binding residues (Zhang, Huang et al. 2000). Zhang et al discussed the possibility of an oxyanion hole formed by backbone nitrogens of the two residues following conserved G101 (with

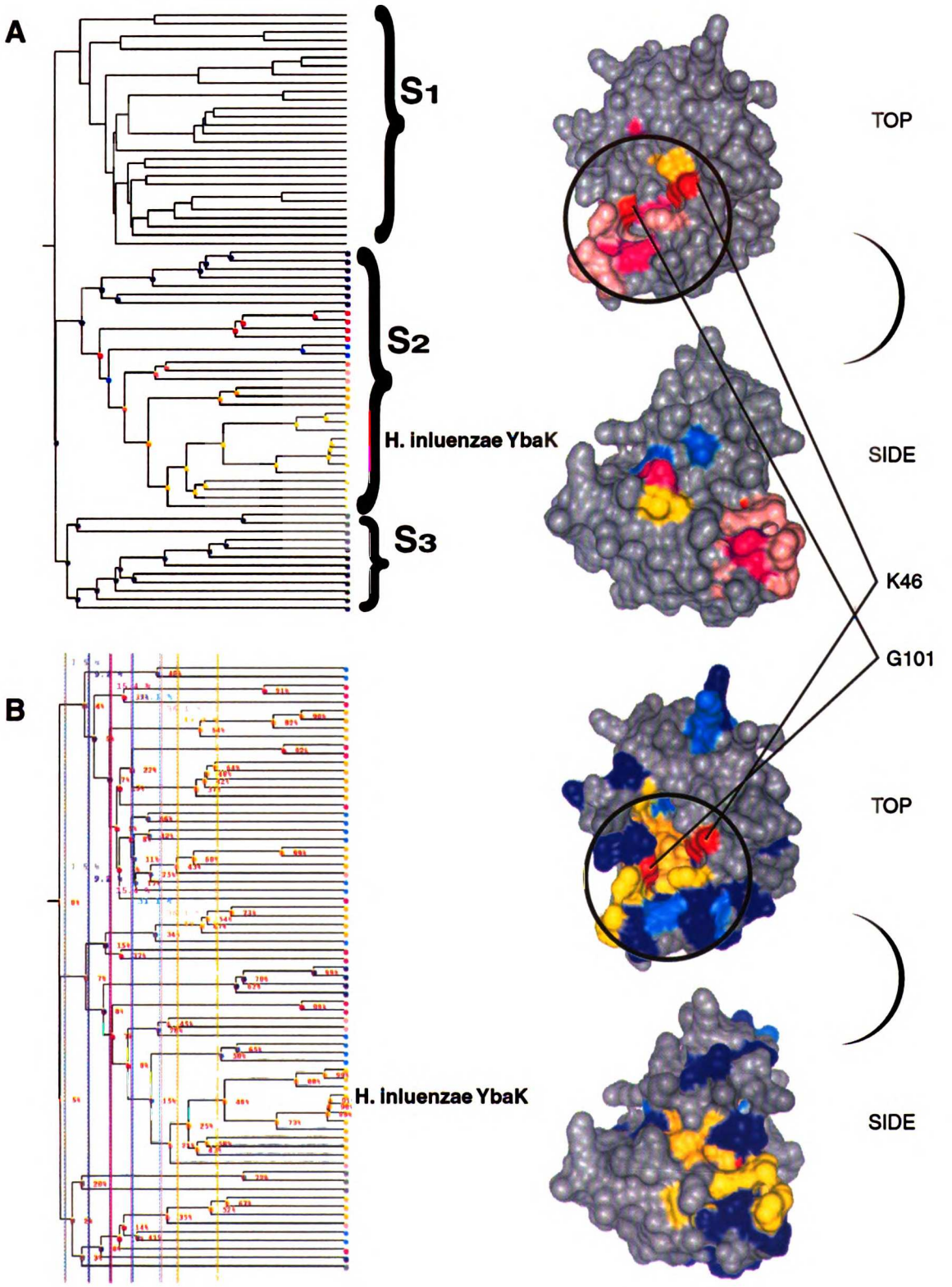
the exception of an arginine in an unknown protein from *Mycobacterium smegmatis* (AAD41809)).

The YbaK family is composed of three large subclades, related by an absolutely conserved lysine, K46. The three subclades are YbaK (S_2 in Figure 28A), an insertion domain in the acceptor stem of prokaryotic prolyl-tRNA synthetases (S_1 in Figure 28A) and a prokaryotic family of hypothetical proteins (S_3 in Figure 28A). Seventy one sequences were used in the JEvTrace analysis. Of these, twenty three formed a distinct subclade containing the *H. influenzae* YbaK sequence.

JEvTrace parent trace analysis, which relies on tracing the conservation of the progenitor sequences of a single selected node (Figure 28A), identified a number of neutral polar and hydrophobic amino acids conserved in the YbaK subclade on the conserved lysine face of YbaK (Figure 28A TOP). This was consistent with the analysis of Zhang et al. Among these conserved positions, JEvTrace identified a cluster of solvent accessible residues above and beyond the proposed oxyanion hole, including Y20, H22, D23, E32 and R132. Together, these residues form a polar surface patch and the wall of the putative binding site. D23 and E32 contribute to the negative face identified by Zhang et al.

Figure 28

JEvTrace analysis of the YbaK protein family. Absolutely conserved positions are colored red. The three major subclades are labeled S_1, S_2 and S_3, and the location of the *H. influenzae* YbaK (PDB:1DBX) in the phylogeny is shown. A represents the results of parent tracing through seven consecutive parent subclades. Color coding corresponds to the colors of the subclades in the phylogeny (**A**). **B** represents the partition trace results for 6 partitions of the phylogeny, ranging from 7% to 50% percent average sequence identity. The per residue score in the partition trace (**B**) corresponds to the color coding as in Figure 27. Black circles in the TOP views of **A** and **B** represent the approximate location of the putative binding site. Graphics of the molecular surfaces were created with Chimera (Huang 1996)(Huang et al) and MSMS (Sanner, Olson et al. 1996) using the SCF format to import JEvTrace results. Graphics of the phylogeny were created with JEvTrace.


A
S1
S2
S3
H. inluenzae YbaK
TOP
SIDE
K46
G101
B
H. inluenzae YbaK
TOP
SIDE

The phylogenetic partition JEvTrace algorithm was performed using 6 partitions from the 7 to 50 average percent sequence identity (Figure 28B). Based on the partition trace algorithm the highest scoring position is S104 (magenta), and then T47, T96, Y98, G102, I103, S129 (orange). Most of these positions are partially shielded from solvent and/or contribute mainchain hydrogen bonding interactions. Eliminating solvent accessible residues left S129, a position which belongs to the neutral polar patch of the putative binding site. Considering residues with less prominent scores (gray, blue, cyan, yellow), the size of the epitope identified by JEvTrace increases considerably and encompasses nearly half of the K46 face (Figure 28B TOP). Together these positions form a partially buried cluster that defines the bottom and walls of the putative ligand binding site spanning across the K46 face. From the fifth level on (Figure 27) all identified positions are on the conserved lysine face of YbaK. Significantly, the loop immediately above the oxyanion hole is disorded in the 1DBX structure. This loop formed by residues 26-30 is not conserved nor does it conserve chemical properties across the phylogeny. However, a number of subclades express invariance at these positions. Due to its proximity to the conserved G101 and one branch of the

J-shaped putative binding site, this disordered region is predicted to contribute to the functional interaction and specificity of YbaK. A number of structural studies by NMR have demonstrated that RNA binding proteins are flexible and undergo conformational changes upon binding (Markus, Hinck et al. 1997; Feng, Tejero et al. 1998; Varani, Gunderson et al. 2000).

Using GRASP (Nicholls 1992) Zhang et al predicted a positively charged patch on the face of the protein opposite K46, and a negatively charged patch on a face adjacent to K46. However, the YbaK structure has the interesting feature of a single conserved lysine separated by a ring of hydrophobic or neutral residues from a circular arrangement of mixed charged residues (Asp, Arg, Glu, Lys). These residues line the perimeter of the K46 face. This is reminiscent of numerous examples of protein-protein interaction, where hydrophobic "rings" of residues are observed to surround polar and charged residues, with the proposed purpose of screening ionic interactions from solvent (Bogan and Thorn 1998; Thorn and Bogan 2001). This potential protein-protein interaction feature of YbaK is additionally supported by evidence that the prolyl-tRNA synthetases (S_1 in Figure 28A) interact with other proteins involved in protein synthesis. There may be additional

surface patches of mixed positively and negatively charged residues in YbaK. However, the positively charged surface identified by Zhang et al. contains the highest number of high scoring positions in the JEvTrace analysis of multiple phylogenetic partitions (Figure 28B). The lysine perimeter patch (Figure 28A,28B TOP) and other potential patches are not conserved nor are they identified completely by the partition algorithm (Figure 28A,28B SIDE), and thus are not expected to be a predominant functional feature of the YbaK family.

JEvTrace analysis suggests that YbaK is involved in a protein-protein interaction requiring a binding site with hydrophobic and polar patches, and an oxyanion hole opposite a conserved lysine. Pursuing the protein-protein interaction hypothesis, it appears that a protruding J-shaped polypeptide volume involving an aspartic or glutamic acid, or a negatively charged cofactor, is a likely ligand for the YbaK binding site. The face opposite this binding site presents a patch of positively charged residues, supporting the hypothesis of a nucleotide binding function for at least some subclades of the YbaK family. Thus although, the YbaK family subtypes only share one conserved amino acid across species, the patterns of subclade sequence conservation suggest a main binding function,

characterized by unique specificity within multiple clades that is spatially centered around the conserved lysine, K46.

Discussion

JEvTrace

The ET method has been a successful tool for analyzing protein functional surfaces using the additional information present in protein phylogenetic trees. However, this approach has been limited by difficulty in producing a dynamic graphical user interface to analyze the data. The optimization involved in producing ET results has previously relied on manually manipulating the underlying data, while certain paths of analysis were inherently inaccessible. Thus, identification of the dominant spatial cluster of invariant residues has been unwieldy. To improve this operational challenge we have constructed JEvTrace, a JAVA suite of algorithms and objects together with a graphical user interface. The algorithms allow the user to identify evolutionary relevant positions based on user selections of subclades (Figure 26A,26B) or partitions of the phylogeny (Figure 28B). This approach introduces new features and parameters in ET analysis. Additional algorithms for tracing subclade conservation through parents or children of a specific subclade (Figure 28A) and subclade conservation comparisons (Figure 26A, 26C) are

available in JEvTrace. The user interface produces interactive graphical results, and access to the underlying sequence, phylogenetic and protein structure data. To perform mapping of ET results and alignment selections to the structural dimension, JEvTrace is dynamically linked to a 3D-structure viewer, Webmol (Walther 1997).

These algorithms (see Methods) allow comparisons of features within a phylogeny in ways that are not directly limited by the topology of the phylogeny, sequence representation bias, or sequence distance and amino acid similarity. The implementation allows an analysis of any possible combinations of subclades within the protein phylogeny. The resulting decompositions of evolutionary sequence data allow multiple definitions of sequence, structure and function homology within a protein family, and hence grant new perspectives to family sequence analysis.

By not considering solvent inaccessible residues, the original ET method relies on protein structures to filter phylogenetic results in order to identify the predicted functional sites. A recognized limitation of the original method was filtering out buried polar side chains within structural clefts (Lichtarge, Bourne et al. 1996). JEvTrace gives access to the entire set of results prior to residue

solvent accessibility filtering. Extensive structural filtering, not limited to solvent accessibility, can be performed in the JAVA structure viewer WebMol (Walther 1997).

JEvTrace facilitates the analysis of other features of protein families. Conserved positions can be found for any subclade in the phylogeny, and the conservation and variability between any set of subclades can be analyzed (subclade comparison Figure 26B, 26C). This functionality can be used to distinguish homologous proteins with different functions, as first suggested by Aloy et al. For a particular subclade, JEvTrace can perform a parent or child trace, identifying the subclade specific conservation within a chain of parent or child subclades of a node (Figure 28A). This method can be used as an ET surrogate if there is lack of significant homology between subclades of a protein family. This was helpful in the analysis of YbaK. JEvTrace can identify the unique residues in a single sequence relative to the considered sequence data. This was useful for our drug design efforts on a malarial cysteine protease (Joachimiak, Chang et al. 2001). JEvTrace also serves as a sequence and structure viewer. Any JEvTrace analysis of the MSA or phylogenetic data can be visualized on available protein structures. This can be useful in the

modeling of protein structure by homology by highlighting the evolutionary contexts for structural analysis within a protein family (Joachimiak, Chang et al. 2001). All of these informatics features address sequence determinants of specificity and similarity using distinct biological data.

In general, the ET approach is more difficult to apply at lower percent sequence identity, owing to problems with building an accurate sequence alignment, especially in theabsence of structural information (Devos and Valencia 2000; Wilson, Kreychman et al. 2000). For example, annotations based on low percent sequence identity pairwise comparisons were a significant source of errors in the initial yeast genome annotation (Mewes, Albermann et al. 1997). Similar alignment problems can occur at the N and C termini of a protein, and even more commonly in loop regions. In addition to requirements of alignment accuracy, ET also has requirements for the minimal amount of sequence information and the related parameter of evolutionary diversity within the protein family. Of the algorithms provided in JEvTrace the parent/child trace and subclade comparisons can be applied with as few as two sequences. The partition trace and subclade trace require more than a pair of subclades, and benefit non-linearly from larger amounts of data. Overall, the two largest effects of

limited sequence data are the signal to noise ratio for the correlation of invariant sites to functional residues and the ability to identify functional specificity and specific functions. As a corollary, until sequence space of a protein family has been sampled sufficiently, insight into the full functions and specificities within a phylogeny remain limited.

MSA Sequence Coloring Format: SCF

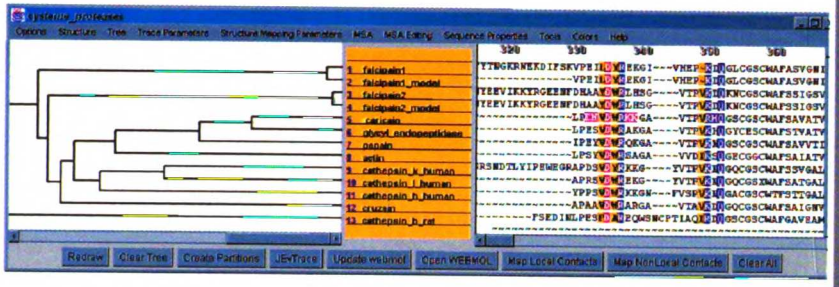
With an ever-increasing set of biological data sources, there is a clear need for sensible standards. We propose SCF: a file format that will encode any user-defined coloring scheme for protein and nucleotide sequences as well as their secondary and tertiary structures, based on the inherent structure of multiple sequence alignments. The format is simple, easy to verify manually, and potentially readable by any alignment or structure viewer.

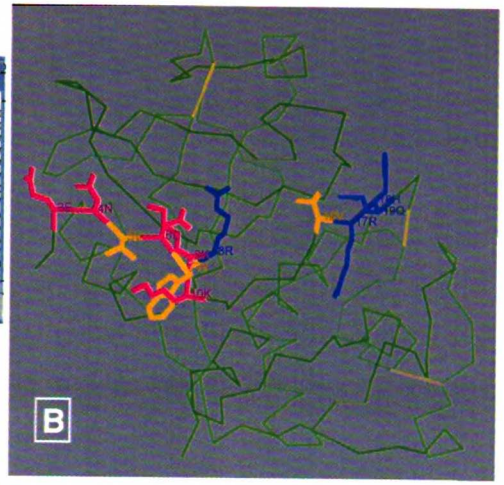
An example protein alignment with a selection of residues including absolutely conserved positions, as well as positions forming two structural epitopes in one of the known protein structures, is shown in Figure 29. It is

Figure 29

An example of the SCF coloring format. **A** represents colored MSA selections. **B** shows the selections mapped to a representative protein structure. **C** is the text encoding of the MSA selections in **A** according to the SCF format specification. The MSA graphics were created with JEvTrace and the structure graphics with WebMol (Walther 1997).

A





C

Position1	Position2	Sequence1	Sequence2	R(RGB)	G(RGB)	B(RGB)	Comment
335	335	0	0	255	200	0	#orange, converved
336	336	0	0	255	0	255	#magenta, conserved
337	337	0	0	255	200	0	#orange, conserved
338	338	0	0	0	0	255	#blue, conserved
350	350	0	0	255	200	0	#orange, gap+conserved
351	351	0	0	0	0	255	#blue, conserved
353	353	0	0	0	0	255	#blue, conserved
333	333	5	5	255	0	255	#magenta epitope
334	334	5	5	255	0	255	#magenta epitope
338	338	5	5	255	0	255	#magenta epitope
339	339	5	5	255	0	255	#magenta epitope
340	340	5	5	0	0	255	#blue epitope
348	348	5	5	0	0	255	#blue epitope
349	349	5	5	0	0	255	#blue epitope
352	352	5	5	0	0	255	#blue epitope

important to note that only selected positions are encoded, a performance and storage asset. Our implementation of an alignment viewer allows transparent interaction with tertiary structures, using the JAVA™ applet WebMol (Walther 1997). In this setting, the color format serves to annotate protein structures with multiple sequence information, and allows comparisons across multiple sequences and structures. This coloring format should aid in the display of results of experiments pertaining to biological sequences and structures. Moreover, it will allow integration of visualized sequence alignment results under a single representation scheme.

Conclusions

We have designed a JAVA™ application, JEvTrace, implementing the Evolutionary Trace method (Lichtarge, Bourne et al. 1996) and its variations. These methods have in common the analysis of protein families through multiple sequence alignments, phylogenetic trees and protein structures. The ET method has proven to be a useful tool for understanding the sequential and structural aspects of protein function, including the analysis of variations relevant to molecular specificity. From an evolutionary perspective, the function of proteins within a protein family encompasses both variation e.g. substrate specificity reflected in amino acids lining substrate binding pockets, and conservation e.g. regions responsible for general enzymatic activity or binding of a common molecular scaffold. While it is trivial to identify absolutely conserved residues, function discovery often requires a context for the predicted or unknown function associated with the absolute conservation pattern.

For the purpose of validating a functional prediction, as in the case of YlxR and YbaK, JEvTrace identified residues clustering around the putative conserved

functional residue(s). These findings supported an RNA-binding prediction for YlxR and most likely a protein-protein interaction interface for YbaK. For functional discovery, as in the case of a new binding epitope in YlxR or the extensive putative binding site and positively charged epitope in YbaK, JEvTrace provided phylogenetic evidence of clusters of residues on the protein surface.

It is hoped that the JEvTrace implementation will lead to analysis of protein families at varying levels of detail, leading to useful decompositions of the data. One of these decompositions comes from the evidence in the evolutionary record of protein sequences. As documented by the biological applications of the ET method, evolutionary data presents evidence allowing the distinction of conserved spatial arrangements of residues versus evolutionary sequence changes with negligible or no effect on function. In unison with experimental data, the decompositions of evolutionary data provided by JEvTrace may enable additional distinctions in the molecular specificity, kinetic and dynamic properties of protein function.

System and Methods

Sequence Family Retrieval and Analysis

We chose two protein structures of unknown function, and retrieved their protein families from the sequence database. The sequence of the structure was used as a query for PSIBLAST (Altschul, Madden et al. 1997) against the GenPept database from NCBI. The sequences were then aligned and phylogenetic trees created with CLUSTALW (Thompson, Higgins et al. 1994) and/or combinations of software from the GCG package (Devereux, Haeberli et al. 1984) including PILEUP (Feng and Doolittle 1987; Higgins and Sharp 1989; Feng and Doolittle 1996) and PAUPSEARCH (Rogers and Swofford 1999).

Algorithm

The binary phylogenetic tree and MSA data are implemented as JAVA™ objects. The phylogenetic tree, assumed to be binary, is modeled as branches and nodes along with an ordering such that each branch shares a node with a parent branch and from zero to two child branches.

Each node in the phylogeny corresponds to a subset of sequences in the MSA. Every phylogenetic branch is represented with an abstract consensus sequence, used to model the corresponding subclade sequence conservation. The implemented algorithm derives a consensus sequence for every subclade of sequences represented in the phylogenetic tree. This information is used to dynamically generate the results based on any user defined subclades or partitions of subclades, by algorithms comparing the appropriate subsets of consensus sequences.

The partition trace variation of the ET method assigns nodes from the tree to a defined partition of the phylogeny. The partition is vertical, meaning perpendicular to the direction of branches in the tree (Figure 25). Sequence conservation in each subclade is compared pairwise to conservation in all other subclades within a given partition. Alternatively, in the subclade trace algorithm, requiring user specification of a set of nodes, the defined nodes are algorithmically treated as a single partition. The subclade trace does not require partitions, and is therefore independent of the topology of the phylogeny. In both algorithms, each position of the MSA is scored by the frequency of conservation of different amino acids in pairs of subclades at that MSA position. In the partition trace,

the score is cumulative across partitions. All scores are normalized if there are more than 7 pairs of invariant subclades at any alignment position. The numerical scores are mapped to a seven color scale (Figure 27), limited by graphical interaction with the structure. Scores can include normalization by the sequence variability of the identified invariant subclades.

JEvTrace also provides the ability to perform a single subclade trace. The user defined subclade is assigned as a parent or child node, and the subclade specific sequence conservation below or above that node is identified. Subclade specific conservation is defined by the set of residues that are conserved in a subclade but not in its parent. The results are a chain of related subclades of the phylogeny, with color coded subclade sequence selections on the MSA and structure. We call this variation of the ET method a parent (or child) trace, and it is especially useful for families with few subclades and cases of highly speciated specificity.

JEvTrace generates results dynamically, displays them on the MSA and enables saving in standard graphics formats or the SCF format. As in the original ET method, absolutely conserved positions are inherently excluded from the analysis. Structural filtering is designated to the WebMol

JAVA™ program, packaged with JEvTrace. Concurrently with WebMol, JEvTrace reads PDB data and aligns the sequence of the structure with a selected sequence in the MSA. This alignment enables JEvTrace to map results and selections from the MSA to the structural dimension. JEvTrace also presents the option of using the Access program (S. Presnell) results to filter the results of the analysis by three states of amino acid solvent accessibility (Defay and Cohen 1996).

Implementation

JEvTrace

The program takes as input an MSA, or an MSA with a corresponding phylogenetic tree. The PILEUP (GCG), CLUSTALW (Thompson, Higgins et al. 1994) and New Hampshire formats (Felsenstein 1989) are recognized. Phylogeny is interpreted as a binary tree with a hypothetical root. Protein structure viewing is designated to the JAVA™ structure viewer WebMol (Walther 1997). Alignment selections in JEvTrace can be mapped to the Chimera structure viewer (Huang 1996) using an earlier version of the SCF format available in JEvTrace. Since many proteins lack representative crystal structures, use of structures in JEvTrace analysis is optional. Currently JEvTrace supports one active WebMol window per session.

Users can select up to seven partitions of the phylogenetic tree, or choose any set of nodes. A number of operations including the ET method can be performed on the selected parts of the phylogeny. The resulting data is independent of structural information and can be viewed and manipulated directly on the MSA of the protein family. To

aid interpretation of the phylogenetic data, the tree can be annotated with the percent sequence identity of all the subclades. The identified positions are visualized on the MSA as well as any available structures (Figure 26,28).

Among the many sequence-structure features of JEvTrace is the ability to highlight the residues in contact with a selected position, based on a residue-residue distance calculation and a distance cutoff. A number of sequence-based features are also available including calculation of alignment position statistics for a variety of physical-chemical properties: molecular volume (Creighton 1992), average pKa (Creighton 1992), hydrogen bonding potential, number of rotatable bonds, hydrophobicity (Karplus 1997).

JEvTrace consists of three graphical canvases: a binary phylogenetic tree, a list of sequence identifiers (e.g. accession codes) and a MSA. The three canvases are aligned by row, such that the terminal nodes (representing individual sequences) of the phylogenetic tree align with their names and amino acid sequences. The tree and alignment canvases are scrollable in two dimensions, and have a practical capacity of more than 150 sequences of less 400 amino acids, on a Pentium workstation with 256M of RAM. All JEvTrace functions are organized into menus and buttons, allowing extensive user interaction with the data.

Any results that are represented graphically in JEvTrace or WebMol can be printed or saved.

Sequence Coloring Format (SCF)

The MSA coloring format exists as a text file with the file extension '.SCF'. It is accurate with respect to the underlying sequence data, given that the sequence(s) remains unchanged in length and order. As a safeguard for the underlying sequence data consistency, the SCF object calculates a MSA checksum variable (see SCF website details). Relational databases and software environments, such as JEvTrace, represent dynamic extensions of the format. The coloring data can exist as an individual file, or can be appended to the actual data file: Multiple Sequence Format (MSF) (GCG) or CLUSTALW (Thompson, Higgins et al. 1994) files in the case of multiple sequence alignment, and a Protein Data Bank (PDB) (Bernstein et al, 1997) file for structural data. Appending the coloring to the underlying data, can allow the transparent annotation by color.

The residue positions of sequences in an alignment are uniquely indexed from top to bottom, using sequence numbers starting at one as rows, and left to right, using alignment

positions starting at zero as columns. The actual file format consists of 6 columns: sequence number, residue number, three columns for the primary color space (Red Green Blue, RGB) designation of the color, and an optional comment/property column (Figure 29C). The last column can accommodate accepted coloring schemes, or can be used to define properties for colors and/or the underlying data. This format accommodates any 24-bit digital color, and allows highlighting of any subset of residues in any subset of sequences of the MSA. The selections are encoded in a hierarchical sorted manner, i.e. smallest to largest sequence position, and within this group smallest to largest sequence, and within those groups, smallest to largest color values.

Our JEvTrace implementation of the SCF format in JAVA™, allows reading MSF, CLUSTALW and PDB data files, and interpretation of the SCF coloring data in each of these contexts. In addition, the underlying sequence data is modeled as JAVA™ objects, whose properties are dynamically updated. In this implementation, it is possible to translate selections between different MSAs sharing at least one sequence. Using a single sequence as a "translator", any selections can be "translated" from one alignment to another, given that both alignments contain

the "translator" sequence. This feature is useful in bridging analysis of families containing distant homologs, performing independent analysis of multiple subclades of a protein family, or updating multiple sequence alignment data.

The JEvTrace and SCF JAVA™ packages have been tested on SGI™ MIPS, Pentium™ Pro (Windows™ and Linux) and Macintosh systems. Both JAVA™ packages are compatible with 1.2 and higher versions of JAVA™.

Acknowledgements

I am deeply grateful for the multiple discussions with Dietlind Gerloff, Dirk Walther, Jonathan Blake, John-Marc Chandonia, Wally Novak and Chern-Sing Goh during the development of the application. Anthony Lau and Elaine Meng provided invaluable comments for the manuscript.

Epilogue

Of special importance in this work has been the overarching theme that effective understanding of protein structure and function in the post-genomic sequence era is related to a subtle interplay between experimental measurements and computational models, analyses and representations. Gene sequences are representations of the complex structure of DNA arranged into genomes. Protein sequences represent a translation of the gene sequence. However, considering primary sequence to tertiary structure correlations or homology to a known structure, the amino acid sequence represents the structure itself. Phylogenies of sequences represent the collection of differences and similarities characteristic of biological evolution. Such representations of complex biological entities and processes facilitate novel computational interpretations of biological data. The combination of biological data and computational methods represents a significant nonadditive advantage in biological analysis and discovery.

Substrate specificity can be changed with even a single residue - however, encoding of a new function or

interaction involves protein surface remodeling and occurs with evolutionary selection of multiple residues. On the organismal level, punctual and gradual aspects of biological evolution can be illustrated with the founder effect. The punctual aspect is represented by the appearance of a new species, the founder. Gradual mutation and selection differentiates the original species into multiple subspecies characterized by unique, distinguishable traits. This phenomenon is an analogy for gene evolution, where genetics and selection lead to multiple homologous genes both within and across genomes.

As has been true in general of knowledge accumulation, increases in the surface area of a field of study are usually unevenly correlated with increases in the depth of the volume knowledge. This phenomenon compounds with the fact that biological systems and their components are as complex as any entities studied by science too date. There is much work remaining before phenotypes can be linked to the molecular biology triad of sequence, structure and function. Computational methods based on models directly relating to experimental data can provide useful insights even in the presence of weak signals within the intrinsically noisy biological information.

Bibliography

Aguejouf, O., E. Malfatti, et al. (2000). "Time related neutralization of two doses acetyl salicylic acid." Thromb Res **100**(4): 317-23.

Aloy, P., E. Querol, et al. (2001). "Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking." J Mol Biol **311**(2): 395-408.

Altschuh, D., A. M. Lesk, et al. (1987). "Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus." J Mol Biol **193**(4): 693-707.

Altschuh, D., T. Vernet, et al. (1988). "Coordinated amino acid changes in homologous protein families." Protein Eng **2**(3): 193-9.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Andrade, M. A., S. I. O'Donoghue, et al. (1998). "Adaptation of protein surfaces to subcellular location." J Mol Biol **276**(2): 517-25.

Anfinsen, C. B. (1973). "Principles that govern the folding of protein chains." Science **181**(96): 223-30.

Anfinsen, C. B., E. Haber, et al. (1961). "The kinetics of formation of native ribonuclease during oxidation of

the reduced polypeptide chain." Proc Natl Acad Sci U S A **47**: 1309-1314.

Astbury, W. T., F. S., et al. (1941). "Nature of the Intramolecular Fold in Alpha-Keratin and Alpha-Myosin." Nature **147**(3736): 696-699.

Aszodi, A., M. J. Gradwell, et al. (1995). "Global fold determination from a small number of distance restraints." J Mol Biol **251**(2): 308-26.

Atwell, S., M. Ultsch, et al. (1997). "Structural plasticity in a remodeled protein-protein interface." Science **278**(5340): 1125-8.

Bae, W., B. Xia, et al. (2000). "Escherichia coli CspA-family RNA chaperones are transcription antiterminators." Proc Natl Acad Sci U S A **97**(14): 7784-9.

Baldwin, E., J. Xu, et al. (1996). "Thermodynamic and structural compensation in "size-switch" core repacking variants of bacteriophage T4 lysozyme." J Mol Biol **259**(3): 542-59.

Baldwin, E. P., O. Hajiseyedjavadi, et al. (1993). "The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme." Science **262**(5140): 1715-8.

Benner, S. A., I. Badcoe, et al. (1994). "Bona fide prediction of aspects of protein conformation. Assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences." J Mol Biol **235**(3): 926-58.

Benner, S. A., D. Gerloff, et al. (1995). "The phospho-beta-galactosidase and synaptotagmin predictions." Proteins **23**(3): 446-53.

Bensing, B. A. and G. M. Dunny (1993). "Cloning and molecular analysis of genes affecting expression of binding substance, the recipient-encoded receptor(s) mediating mating aggregate formation in Enterococcus faecalis." J Bacteriol **175**(22): 7421-9.

Berger, A. and I. Schechter (1970). "Mapping the active site of papain with the aid of peptide substrates and inhibitors." Philos Trans R Soc Lond B Biol Sci **257**(813): 249-64.

Bernstein, F. C., T. F. Koetzle, et al. (1978). "The protein data bank: a computer-based archival file for macromolecular structures." Arch Biochem Biophys **185**(2): 584-91.

Blaber, M., X. J. Zhang, et al. (1993). "Structural basis of amino acid alpha helix propensity." Science **260**(5114): 1637-40.

Blaney, J. M., G. M. Crippen, et al. (1984-1994). DGEOM, Copyright 1990 DuPont Corporation. Copyright 1995 Chiron Corporation.

Blundell, T. L. (1991). "Comparative analysis of protein three-dimensional structures and an approach to the inverse folding problem." Ciba Found Symp **161**: 28-36; discussion 37-51.

Bogan, A. A. and K. S. Thorn (1998). "Anatomy of hot spots in protein interfaces." J Mol Biol **280**(1): 1-9.

Bossard, M. J., T. A. Tomaszek, et al. (1996). "Proteolytic activity of human osteoclast cathepsin K. Expression, purification, activation, and substrate identification." J Biol Chem **271**(21): 12517-24.

Bower, M. J., F. E. Cohen, et al. (1997). "Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool." J Mol Biol **267**(5): 1268-82.

Bowie, J. U., R. Luthy, et al. (1991). "A method to identify protein sequences that fold into a known three-dimensional structure." Science **253**(5016): 164-70.

Brenner, S. E. (1999). "Errors in genome annotation." Trends Genet **15**(4): 132-3.

Burns, D. M. and I. R. Beacham (1986). "Identification and sequence analysis of a silent gene (ushA0) in Salmonella typhimurium." J Mol Biol **192**(2): 163-75.

Bystroff, C., K. T. Simons, et al. (1996). "Local sequence-structure correlations in proteins." Curr Opin Biotechnol **7**(4): 417-21.

Bystroff, C., V. Thorsson, et al. (2000). "HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins." J Mol Biol **301**(1): 173-90.

Casari, G., C. Sander, et al. (1995). "A method to predict functional residues in proteins." Nat Struct Biol **2**(2): 171-8.

Chandonia, J. M. and M. Karplus (1999). "New methods for accurate prediction of protein secondary structure." Proteins **35**(3): 293-306.

Chelvanayagam, G., A. Eggenschwiler, et al. (1997). "An analysis of simultaneous variation in protein structures." Protein Eng **10**(4): 307-16.

Chelvanayagam, G., L. Knecht, et al. (1998). "A combinatorial distance-constraint approach to predicting protein tertiary models from known secondary structure." Folding & Design **3**(3): 149-60.

Cho, S. J., M. G. Lee, et al. (2000). "Crystal structure of Escherichia coli CyaY protein reveals a previously unidentified fold for the evolutionarily conserved

frataxin family." Proc Natl Acad Sci U S A **97**(16): 8932-7.

Chothia, C. and A. M. Lesk (1986). "The relation between the divergence of sequence and structure in proteins." Embo J **5**(4): 823-6.

Chothia, C. and A. M. Lesk (1987). "The evolution of protein structures." Cold Spring Harb Symp Quant Biol **52**: 399-405.

Cohen, F. E., M. J. Sternberg, et al. (1982). "Analysis and prediction of the packing of alpha-helices against a beta-sheet in the tertiary structure of globular proteins." J Mol Biol **156**(4): 821-62.

Creighton, T. E. (1992). Proteins: Structures and Molecular Properties. New York, W.H. Freeman.

Crippen, G. M. and V. N. Viswanadhan (1984). "A potential function for conformational analysis of proteins." Int J Pept Protein Res **24**(3): 279-96.

Dayhoff, M. O. (1972). Atlas of Protein Sequence and Structure, National Biomedical Research Foundation.

Defay, T. R. and F. E. Cohen (1996). "Multiple sequence information for threading algorithms." J Mol Biol **262**(2): 314-23.

Devauchelle, C., A. Grossmann, et al. (2001). "Rate matrices for analyzing large families of protein sequences." J Comput Biol **8**(4): 381-99.

Devereux, J., P. Haeberli, et al. (1984). "A comprehensive set of sequence analysis programs for the VAX." Nucleic Acids Res **12**(1 Pt 1): 387-95.

Devos, D. and A. Valencia (2000). "Practical limits of function prediction." Proteins **41**(1): 98-107.

Di Gennaro, J. A., N. Siew, et al. (2001). "Enhanced functional annotation of protein sequences via the use of structural descriptors." J Struct Biol **134**(2-3): 232-45.

dos Remedios, C. G. and D. D. Thomas (2001). "An overview of actin structure and actin-binding proteins." Results Probl Cell Differ **32**: 1-7.

Dosztanyi, Z., A. Fiser, et al. (1997). "Stabilization centers in proteins: identification, characterization and predictions." J Mol Biol **272**(4): 597-612.

Druker, B. J., M. Talpaz, et al. (2001). "Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia." N Engl J Med **344**(14): 1031-7.

Du, P. and I. Alkorta (1994). "Sequence divergence analysis for the prediction of seven-helix membrane protein structures: I. Comparison with bacteriorhodopsin." Protein Eng **7**(10): 1221-9.

Edsall, J. T., P. J. Flory, et al. (1966). "A proposal of standard conventions and nomenclature for the description of polypeptide conformations." J Mol Biol **15**(1): 399-407.

Eggleson, K. K., K. L. Duffin, et al. (1999). "Identification and characterization of falcilysin, a metallopeptidase involved in hemoglobin catabolism within the malaria parasite Plasmodium falciparum." J Biol Chem **274**(45): 32411-7.

Emanueli, C., E. F. Grady, et al. (1998). "Acute ACE inhibition causes plasma extravasation in mice that is mediated by bradykinin and substance P." Hypertension **31**(6): 1299-304.

Engel, C. (2002). Wild Health: how animals keep themselves healthy and what we can learn from them. Boston, Houghton Mifflin Company.

Epstain, C.J., R. F. Goldberger, et al. (1963). Cold Spring Harbor Symp. Quant. Biol., Cold Spring Harbor.

Ewing, T. J., S. Makino, et al. (2001). "DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases." J Comput Aided Mol Des **15**(5): 411-28.

Falicov, A. and F. E. Cohen (1996). "A surface of minimum area metric for the structural comparison of proteins." J Mol Biol **258**(5): 871-92.

Fariselli, P. and R. Casadio (1999). "A neural network based predictor of residue contacts in proteins." Protein Eng **12**(1): 15-21.

Fariselli, P. and R. Casadio (2000). "Prediction of the number of residue contacts in proteins." Proc Int Conf Intell Syst Mol Biol **8**: 146-51.

Fariselli, P. and R. Casadio (2001). "Prediction of disulfide connectivity in proteins." Bioinformatics **17**(10): 957-64.

Fariselli, P. and R. Casadio (2001). "RCNPRED: prediction of the residue co-ordination numbers in proteins." Bioinformatics **17**(2): 202-4.

Fariselli, P., O. Olmea, et al. (2001). "Prediction of contact maps with neural networks and correlated mutations." Protein Eng **14**(11): 835-43.

Felsenstein, J. (1989). "PHYLIP -- Phylogeny Phylogeny Inference Package (Version 3.2)." Cladistics **5**: 164-166.

Feng, D. F. and R. F. Doolittle (1987). "Progressive sequence alignment as a prerequisite to correct phylogenetic trees." J Mol Evol **25**(4): 351-60.

Feng, D. F. and R. F. Doolittle (1996). "Progressive alignment of amino acid sequences and construction of phylogenetic trees from them." Methods Enzymol **266**: 368-82.

Feng, W., R. Tejero, et al. (1998). "Solution NMR structure and backbone dynamics of the major cold-shock protein (CspA) from Escherichia coli: evidence for conformational dynamics in the single-stranded RNA-binding site." Biochemistry **37**(31): 10881-96.

Fisher, N. D. and N. K. Hollenberg (2001). "Is there a future for renin inhibitors?" Expert Opin Investig Drugs **10**(3): 417-26.

Flores, T. P., C. A. Orengo, et al. (1993). "Comparison of conformational characteristics in structurally similar protein pairs." Protein Sci **2**(11): 1811-26.

Foster, P. G. and D. A. Hickey (1999). "Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions." J Mol Evol **48**(3): 284-90.

Francis, S. E., I. Y. Gluzman, et al. (1996). "Characterization of native falcipain, an enzyme involved in Plasmodium falciparum hemoglobin degradation." Mol Biochem Parasitol **83**(2): 189-200.

Garvin, L. D. and S. C. Hardies (1991). "Temporal and topological clustering of diverged residues among enterobacterial dihydrofolate reductases." Mol Biol Evol **8**(5): 654-68.

Gaucher, E. A., M. M. Miyamoto, et al. (2001). "Function-structure analysis of proteins using covarion-based

evolutionary approaches: Elongation factors." Proc Natl Acad Sci U S A **98**(2): 548-52.

Gerloff, D. L., G. M. Cannarozzi, et al. (1999). "Evolutionary, mechanistic, and predictive analyses of the hydroxymethyldihydropterin pyrophosphokinase family of proteins." Biochem Biophys Res Commun **254**(1): 70-6.

Gerloff, D. L., M. Joachimiak, et al. (1998). "Structure prediction in a post-genomic environment: a secondary and tertiary structural model for the initiation factor 5A family." Biochem Biophys Res Commun **251**(1): 173-81.

Gerlt, J. A. and P. C. Babbitt (2000). "Can sequence determine function?" Genome Biol **1**(5): REVIEWS0005.

Gerstein, M., A. M. Lesk, et al. (1994). "Structural mechanisms for domain movements in proteins." Biochemistry **33**(22): 6739-49.

Gerstein, M., E. L. Sonnhammer, et al. (1994). "Volume changes in protein evolution." J Mol Biol **236**(4): 1067-78.

Gobel, U., C. Sander, et al. (1994). "Correlated mutations and residue contacts in proteins." Proteins **18**(4): 309-17.

Godzik, A., A. Kolinski, et al. (1992). "Topology fingerprint approach to the inverse protein folding problem." J Mol Biol **227**(1): 227-38.

Godzik, A., J. Skolnick, et al. (1993). "Regularities in interaction patterns of globular proteins." Protein Eng **6**(8): 801-10.

Goh, C. S., A. A. Bogan, et al. (2000). "Co-evolution of proteins with their interaction partners." J Mol Biol **299**(2): 283-93.

Gorre, M. E., M. Mohammed, et al. (2001). "Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification." Science **293**(5531): 876-80.

Gouldson, P. R., C. Higgs, et al. (2000). "Dimerization and domain swapping in G-protein-coupled receptors: a computational study." Neuropsychopharmacology **23**(4 Suppl): S60-77.

Grill, S., I. Moll, et al. (2001). "Modulation of ribosomal recruitment to 5'-terminal start codons by translation initiation factors IF2 and IF3." FEBS Lett **495**(3): 167-71.

Grishin, N. V. and M. A. Phillips (1994). "The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences." Protein Sci **3**(12): 2455-8.

Gromiha, M. M. and S. Selvaraj (2001). "Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction." J Mol Biol **310**(1): 27-32.

Gu, X. (1999). "Statistical methods for testing functional divergence after gene duplication." Mol Biol Evol **16**(12): 1664-74.

Gu, X. (2001). "Mathematical modeling for functional divergence after gene duplication." J Comput Biol **8**(3): 221-34.

Gu, X. (2001). "A site-specific measure for rate difference after gene duplication or speciation." Mol Biol Evol **18**(12): 2327-30.

Han, K. F. and D. Baker (1995). "Recurring local sequence motifs in proteins." J Mol Biol **251**(1): 176-87.

Han, K. F. and D. Baker (1996). "Global properties of the mapping between local amino acid sequence and local structure in proteins." Proc Natl Acad Sci U S A **93**(12): 5814-8.

Han, K. F., C. Bystroff, et al. (1997). "Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns." Protein Sci **6**(7): 1587-90.

Harrison, P. M. (1996). Analysis and prediction of protein structure: Disulphide bridges. London, UK, University College, London: 95-110.

Harrison, P. M., H. S. Chan, et al. (1999). "Thermodynamics of model prions and its implications for the problem of prion protein folding." J Mol Biol **286**(2): 593-606.

Harrison, P. M., H. S. Chan, et al. (2001). "Conformational propagation with prion-like characteristics in a simple model of protein folding." Protein Sci **10**(4): 819-35.

Henikoff, S. and J. G. Henikoff (1993). "Performance evaluation of amino acid substitution matrices." Proteins **17**(1): 49-61.

Higgins, D. G. and P. M. Sharp (1989). "Fast and sensitive multiple sequence alignments on a microcomputer." Comput Appl Biosci **5**(2): 151-3.

Hill, J., L. Tyas, et al. (1994). "High level expression and characterisation of Plasmepsin II, an aspartic proteinase from Plasmodium falciparum." FEBS Lett **352**(2): 155-8.

Hishigaki, H., K. Nakai, et al. (2001). "Assessment of prediction accuracy of protein function from protein--protein interaction data." Yeast **18**(6): 523-31.

Hoh, J. H. (1998). "Functional protein domains from the thermally driven motion of polypeptide chains: a proposal." Proteins **32**(2): 223-8.

Holmbeck, S. M., M. P. Foster, et al. (1998). "High-resolution solution structure of the retinoid X receptor DNA-binding domain." J Mol Biol **281**(2): 271-84.

Huang, C. C., G. S. Couch, et al. (1996). "Chimera: An Extensible Molecular Modeling Application Constructed Using Standard Components." Pac. Symp. Biocomput. **1**: 724.

Huang, C. C., Couch, G.S., Pettersen, E.F., and Ferrin, T.E (1996). "Chimera: An Extensible Molecular Modeling Application Constructed Using Standard Components." Pacific Symposium on Biocomputing **1**: 724.

Huggins, M. L. (1943). "The Structure of Fibrous Proteins." Chem. Rev.(32): 195-218.

Hutchinson, E. G., R. B. Sessions, et al. (1998). "Determinants of strand register in antiparallel beta-sheets of proteins." Protein Sci **7**(11): 2287-300.

Innis, C. A., J. Shi, et al. (2000). "Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis." Protein Eng **13**(12): 839-47.

Joachimiak, M. P., C. Chang, et al. (2001). "The Impact of Whole Genome Sequence Data on Drug Discovery--A Malaria Case Study." Mol Med **7**(10): 698-710.

Johnson, J. M. and G. M. Church (2000). "Predicting ligand-binding function in families of bacterial receptors." Proc Natl Acad Sci U S A **97**(8): 3965-70.

Jones, D. T., M. Tress, et al. (1999). "Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure." Proteins **Suppl 3**: 104-11.

Karplus, P. A. (1997). "Hydrophobicity regained." Protein Sci **6**(6): 1302-7.

Kimura, M. (1968). "Evolutionary rate at the molecular level." Nature **217**(129): 624-6.

Kitazoe, Y., Y. Kurihara, et al. (2001). "A new theory of phylogeny inference through construction of multidimensional vector space." Mol Biol Evol **18**(5): 812-28.

Kolesov, G., H. W. Mewes, et al. (2001). "SNAPping up functionally related genes based on context information: a colinearity-free approach." J Mol Biol **311**(4): 639-56.

Korber, B. T., R. M. Farber, et al. (1993). "Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis." Proc Natl Acad Sci U S A **90**(15): 7176-80.

Lambros, C. and J. P. Vanderberg (1979). "Synchronization of Plasmodium falciparum erythrocytic stages in culture." J Parasitol. **65**(3): 418-20.

Lamperti, E. D., J. M. Kittelberger, et al. (1992). "Corruption of genomic databases with anomalous sequence." Nucleic Acids Res **20**(11): 2741-7.

Landgraf, R., D. Fischer, et al. (1999). "Analysis of heregulin symmetry by weighted evolutionary tracing." Protein Eng **12**(11): 943-51.

Le Bonniec, S., C. Deregnaucourt, et al. (1999). "Plasmepsin II, an acidic hemoglobinase from the Plasmodium falciparum food vacuole, is active at neutral pH on the host erythrocyte membrane skeleton." J Biol Chem **274**(20): 14218-23.

Li, R., X. Chen, et al. (1996). "Structure-based design of parasitic protease inhibitors." Bioorg Med Chem **4**(9): 1421-7.

Li, W. H., C. I. Wu, et al. (1985). "A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes." Mol Biol Evol **2**(2): 150-74.

Li, Z., X. Chen, et al. (1994). "Anti-malarial drug development using models of enzyme structure." Chem Biol **1**(1): 31-7.

Liberles, D. A., D. R. Schreiber, et al. (2001). "The adaptive evolution database (TAED)." Genome Biol **2**(8): RESEARCH0028.

Lichtarge, O., H. R. Bourne, et al. (1996). "Evolutionarily conserved Galphabetagamma binding surfaces support a model of the G protein-receptor complex." Proc Natl Acad Sci U S A **93**(15): 7507-11.

Lichtarge, O., H. R. Bourne, et al. (1996). "An evolutionary trace method defines binding surfaces common to protein families." J Mol Biol **257**(2): 342-58.

Lichtarge, O., M. E. Sowa, et al. (2002). "Evolutionary traces of functional surfaces along G protein signaling pathway." Methods Enzymol **344**: 536-56.

Lichtarge, O., K. R. Yamamoto, et al. (1997). "Identification of functional surfaces of the zinc binding domains of intracellular receptors." J Mol Biol **274**(3): 325-37.

Liu, Y., D. Zhao, et al. (1992). "A systematic comparison of three structure determination methods from NMR data: dependence upon quality and quantity of data." J Biomol NMR **2**(4): 373-88.

Lo Conte, L., S. E. Brenner, et al. (2002). "SCOP database in 2002: refinements accommodate structural genomics." Nucleic Acids Res **30**(1): 264-7.

Loll, P. J., D. Picot, et al. (1995). "The structural basis of aspirin activity inferred from the crystal structure of inactivated prostaglandin H2 synthase." Nat Struct Biol **2**(8): 637-43.

Lowman, H. B., S. H. Bass, et al. (1991). "Selecting high-affinity binding proteins by monovalent phage display." Biochemistry **30**(45): 10832-8.

Lund, O., J. Hansen, et al. (1996). "Relationship between protein structure and geometrical constraints." Protein Sci **5**(11): 2217-25.

Maiorov, V. N. and G. M. Crippen (1992). "Contact potential that recognizes the correct folding of globular proteins." J Mol Biol **227**(3): 876-88.

Marchler-Bauer, A. and S. H. Bryant (1997). "Measures of threading specificity and accuracy." Proteins **Suppl 1**: 74-82.

Marcotte, E. M., M. Pellegrini, et al. (1999). "A combined algorithm for genome-wide prediction of protein function." Nature **402**(6757): 83-6.

Mark, K. S. and T. P. Davis (2000). "Stroke: development, prevention and treatment with peptidase inhibitors." Peptides **21**(12): 1965-73.

Markus, M. A., A. P. Hinck, et al. (1997). "High resolution solution structure of ribosomal protein L11-C76, a helical protein with a flexible loop that becomes structured upon binding to RNA." Nat Struct Biol **4**(1): 70-7.

Matthews, B. W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." Biochim Biophys Acta **405**(2): 442-51.

McGrath, M. E. (1999). "The lysosomal cysteine proteases." Annu Rev Biophys Biomol Struct **28**: 181-204.

Messier, W. and C. B. Stewart (1997). "Episodic adaptive evolution of primate lysozymes." Nature **385**(6612): 151-4.

Mewes, H. W., K. Albermann, et al. (1997). "MIPS: a database for protein sequences, homology data and yeast genome information." Nucleic Acids Res **25**(1): 28-30.

Mewes, H. W., D. Frishman, et al. (2002). "MIPS: a database for genomes and protein sequences." Nucleic Acids Res **30**(1): 31-4.

Michnick, S. W. and E. Shakhnovich (1998). "A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies." Fold Des **3**(4): 239-51.

Mirny, L. and E. Shakhnovich (2001). "Evolutionary conservation of the folding nucleus." J Mol Biol **308**(2): 123-9.

Miyazawa, S. and R. L. Jernigan (1985). "Estimation of Effective Interresidue Contact Energies from Protein

Crystal Structures: Quasi-Chemical Approximation." Macromolecules(18): 534-552.

Miyazawa, S. and R. L. Jernigan (1993). "A new substitution matrix for protein sequence searches based on contact frequencies in protein structures." Protein Eng **6**(3): 267-78.

Morshauser, R. C., W. Hu, et al. (1999). "High-resolution solution structure of the 18 kDa substrate-binding domain of the mammalian chaperone protein Hsc70." J Mol Biol **289**(5): 1387-403.

Muskett, F. W., T. A. Frenkiel, et al. (1998). "High resolution structure of the N-terminal domain of tissue inhibitor of metalloproteinases-2 and characterization of its interaction site with matrix metalloproteinase-3." J Biol Chem **273**(34): 21736-43.

Neher, E. (1994). "How frequent are correlated changes in families of protein sequences?" Proc Natl Acad Sci U S A **91**(1): 98-102.

Nicholls, A. (1992). GRASP: graphical representation and analysis of surface properties. New York, Columbia University.

Nilges, M. (1995). "Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities." J Mol Biol **245**(5): 645-60.

Nilges, M. (1997). "Ambiguous distance data in the calculation of NMR structures." Fold Des **2**(4): S53-7.

Nilges, M., M. J. Macias, et al. (1997). "Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin." J Mol Biol **269**(3): 408-22.

Olmea, O., B. Rost, et al. (1999). "Effective use of sequence correlation and conservation in fold recognition." J Mol Biol **293**(5): 1221-39.

Olmea, O. and A. Valencia (1997). "Improving contact predictions by the combination of correlated mutations and other sources of sequence information." Fold Des **2**(3): S25-32.

Orengo, C. A., J. E. Bray, et al. (2002). "The CATH protein family database: A resource for structural and functional annotation of genomes." Proteomics **2**(1): 11-21.

Orengo, C. A., J. E. Bray, et al. (1999). "Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction." Proteins **37**(S3): 149-170.

Osipiuk, J., P. Gornicki, et al. (2001). "Streptococcus pneumonia YlxR at 1.35 A shows a putative new fold." Acta Crystallogr D Biol Crystallogr **57**(Pt 11): 1747-51.

Otzen, D. E. and A. R. Fersht (1995). "Side-chain determinants of beta-sheet stability." Biochemistry **34**(17): 5718-24.

Panchenko, A., A. Marchler-Bauer, et al. (1999). "Threading with explicit models for evolutionary conservation of structure and sequence." Proteins **37**(S3): 133-140.

Pauling, L. (1974). "Molecular basis of biological specificity." Nature **248**(451): 769-71.

Pauling, L. and R. B. Corey (1951). "Configurations of Polypeptide Chains with Favored Orientations Around Single Bonds: Two New Pleated Sheets." Proc Natl Acad Sci U S A **37**: 729-740.

Pauling, L., R. B. Corey, et al. (1951). "The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain." Proc Natl Acad Sci U S A **37**: 205-211.

Pazos, F., M. Helmer-Citterich, et al. (1997). "Correlated mutations contain information about protein-protein interaction." J Mol Biol **271**(4): 511-23.

Pazos, F. and A. Valencia (2001). "Similarity of phylogenetic trees as indicator of protein-protein interaction." Protein Eng **14**(9): 609-14.

Pearl, F., A. E. Todd, et al. (2000). "Using the CATH domain database to assign structures and functions to the genome sequences." Biochem Soc Trans **28**(2): 269-75.

Pieper, U., N. Eswar, et al. (2002). "MODBASE, a database of annotated comparative protein structure models." Nucleic Acids Res **30**(1): 255-9.

Plaxco, K. W., K. T. Simons, et al. (1998). "Contact order, transition state placement and the refolding rates of single domain proteins." J Mol Biol **277**(4): 985-94.

Pollock, D. D. and W. R. Taylor (1997). "Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution." Protein Eng **10**(6): 647-57.

Pollock, D. D., W. R. Taylor, et al. (1999). "Coevolving protein residues: maximum likelihood identification and relationship to structure." J Mol Biol **287**(1): 187-98.

Poteete, A. R., D. P. Sun, et al. (1991). "Second-site revertants of an inactive T4 lysozyme mutant restore activity by restructuring the active site cleft." Biochemistry **30**(5): 1425-32.

Pritchard, L. and M. J. Dufton (1999). "Evolutionary trace analysis of the Kunitz/BPTI family of proteins: functional divergence may have been based on conformational adjustment." J Mol Biol **285**(4): 1589-607.

Ptitsyn, O. B. (1998). "Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes?" J Mol Biol **278**(3): 655-66.

Ptitsyn, O. B. and K. L. Ting (1999). "Non-functional conserved residues in globins and their possible role as a folding nucleus." J Mol Biol **291**(3): 671-82.

Raillard, S., A. Krebber, et al. (2001). "Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes." Chem Biol **8**(9): 891-8.

Reese, M. G., O. Lund, et al. (1996). "Distance distributions in proteins: a six-parameter representation." Protein Eng **9**(9): 733-40.

Ring, C. S., E. Sun, et al. (1993). "Structure-based inhibitor design by using protein models for the development of antiparasitic agents." Proc Natl Acad Sci U S A **90**(8): 3583-7.

Rodionov, M. A. and M. S. Johnson (1994). "Residue-residue contact substitution probabilities derived from aligned three-dimensional structures and the identification of common folds." Protein Sci **3**(12): 2366-77.

Rogers, J. S. and D. L. Swofford (1999). "Multiple local maxima for likelihoods of phylogenetic trees: a simulation study." Mol Biol Evol **16**(8): 1079-85.

Rosenthal, P. J., K. Kim, et al. (1987). "Identification of three stage-specific proteinases of Plasmodium falciparum." J Exp Med **166**(3): 816-21.

Rosenthal, P. J., J. H. McKerrow, et al. (1988). "A malarial cysteine proteinase is necessary for hemoglobin degradation by Plasmodium falciparum." J Clin Invest **82**(5): 1560-6.

Rosenthal, P. J. and R. G. Nelson (1992). "Isolation and characterization of a cysteine proteinase gene of Plasmodium falciparum." Mol Biochem Parasitol **51**(1): 143-52.

Rosenthal, P. J., J. E. Olson, et al. (1996). "Antimalarial effects of vinyl sulfone cysteine proteinase inhibitors." Antimicrob Agents Chemother **40**(7): 1600-3.

Rosenthal, P. J., W. S. Wollish, et al. (1991). "Antimalarial effects of peptide inhibitors of a Plasmodium falciparum cysteine proteinase." J Clin Invest **88**(5): 1467-72.

Rost, B. and C. Sander (1994). "Combining evolutionary information and neural networks to predict protein secondary structure." Proteins **19**(1): 55-72.

Russell, R. B. and G. J. Barton (1994). "Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility." J Mol Biol **244**(3): 332-50.

Russell, R. B., P. D. Sasieni, et al. (1998). "Supersites within superfolds. Binding site similarity in the absence of homology." J Mol Biol **282**(4): 903-18.

Sanchez, R. and A. Sali (1997). "Evaluation of comparative protein structure modeling by MODELLER-3." Proteins **Suppl 1**: 50-8.

Sanner, M. F., A. J. Olson, et al. (1995). Fast and robust computation of molecular surfaces. Proc. 11th ACM Symp. Comp. Geom.

Sanner, M. F., A. J. Olson, et al. (1996). "Reduced surface: an efficient way to compute molecular surfaces." Biopolymers **38**(3): 305-20.

Selbig, J. and P. Argos (1998). "Relationships between protein sequence and structure patterns based on residue contacts." Proteins **31**(2): 172-85.

Shakhnovich, E., V. Abkevich, et al. (1996). "Conserved residues and the mechanism of protein folding." Nature **379**(6560): 96-8.

Shenai, B. R., P. S. Sijwali, et al. (2000). "Characterization of native and recombinant falcipain-2, a principal trophozoite cysteine protease and essential hemoglobinase of Plasmodium falciparum." J Biol Chem **275**(37): 29000-10.

Shindyalov, I. N., N. A. Kolchanov, et al. (1994). "Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?" Protein Eng **7**(3): 349-58.

Sieber, V., C. A. Martinez, et al. (2001). "Libraries of hybrid proteins from distantly related sequences." Nat Biotechnol **19**(5): 456-60.

Simons, K. T., C. Kooperberg, et al. (1997). "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions." J Mol Biol **268**(1): 209-25.

Singer, G. A. and D. A. Hickey (2000). "Nucleotide bias causes a genomewide bias in the amino acid composition of proteins." Mol Biol Evol **17**(11): 1581-8.

Singer, M. S., L. Oliveira, et al. (1995). "Potential ligand-binding residues in rat olfactory receptors identified by correlated mutation analysis." Receptors Channels **3**(2): 89-95.

Sippl, M. J. (1990). "Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins." J Mol Biol **213**(4): 859-83.

Smith, R. F. and T. F. Smith (1992). "Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling." Protein Eng **5**(1): 35-41.

Sowa, M. E., W. He, et al. (2001). "Prediction and confirmation of a site critical for effector regulation of RGS domain activity." Nat Struct Biol **8**(3): 234-7.

Sowa, M. E., W. He, et al. (2000). "A regulator of G protein signaling interaction surface linked to effector specificity." Proc Natl Acad Sci U S A **97**(4): 1483-8.

Stevens, R. C., S. Yokoyama, et al. (2001). "Global efforts in structural genomics." Science **294**(5540): 89-92.

Stricker, E. M. and A. F. Sved (2000). "Thirst." Nutrition **16**(10): 821-6.

Strickland, L., G. D. Letson, et al. (2001). "Gastrointestinal stromal tumors." Cancer Control **8**(3): 252-61.

Strimmer, K. and V. Moulton (2000). "Likelihood analysis of phylogenetic networks using directed graphical models." Mol Biol Evol **17**(6): 875-81.

Swerdlow, J. L. and L. Johnson (2000). Nature's Medicines: plants that heal, National Geographic Society.

Tatusov, R. L., D. A. Natale, et al. (2001). "The COG database: new developments in phylogenetic classification of proteins from complete genomes." Nucleic Acids Res **29**(1): 22-8.

Taverna, D. M. and R. A. Goldstein (2002). "Why are proteins so robust to site mutations?" J Mol Biol **315**(3): 479-84.

Taylor, W. R. (1997). "Residual colours: a proposal for aminochromography." Protein Eng **10**(7): 743-6.

Tebas, P. and W. G. Powderly (2000). "Nelfinavir mesylate." Expert Opin Pharmacother **1**(7): 1429-40.

Thomas, D. J., G. Casari, et al. (1996). "The prediction of protein contacts from multiple sequence alignments." Protein Eng **9**(11): 941-8.

Thomas, P. D. and K. A. Dill (1996). "Statistical potentials extracted from protein structures: how accurate are they?" J Mol Biol **257**(2): 457-69.

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

Thompson, M. J. and R. A. Goldstein (1997). "Predicting protein secondary structure with probabilistic schemata of evolutionarily derived information." Protein Sci **6**(9): 1963-75.

Thorn, K. S. and A. A. Bogan (2001). "ASEdb: a database of alanine mutations and their effects on the free energy

of binding in protein interactions." Bioinformatics **17**(3): 284-5.

Turk, D., G. Guncar, et al. (1998). "Revised definition of substrate binding sites of papain-like cysteine proteases." Biol Chem **379**(2): 137-47.

Tuveson, D. A., N. A. Willis, et al. (2001). "STI571 inactivation of the gastrointestinal stromal tumor c-KIT oncoprotein: biological and clinical implications." Oncogene **20**(36): 5054-8.

Varani, L., S. I. Gunderson, et al. (2000). "The NMR structure of the 38 kDa U1A protein - PIE RNA complex reveals the basis of cooperativity in regulation of polyadenylation by human U1A protein." Nat Struct Biol **7**(4): 329-35.

Vendruscolo, M., E. Kussell, et al. (1997). "Recovery of protein structure from contact maps." Fold Des **2**(5): 295-306.

Vetter, I. R., W. A. Baase, et al. (1996). "Protein structural plasticity exemplified by insertion and deletion mutants in T4 lysozyme." Protein Sci **5**(12): 2399-415.

Wagele, J. W. and F. Rodding (1998). "A priori estimation of phylogenetic information conserved in aligned sequences." Mol Phylogenet Evol **9**(3): 358-65.

Walther, D. (1997). "WebMol--a Java-based PDB viewer." Trends Biochem Sci **22**(7): 274-5.

Walther, D. and F. E. Cohen (1999). "Conformational attractors on the Ramachandran map." Acta Crystallogr D Biol Crystallogr **55 (Pt 2)**: 506-17.

Weber, J. L. (1987). "Analysis of sequences from the extremely A + T-rich genome of Plasmodium falciparum." Gene **52**(1): 103-9.

Wells, J. A. (1991). "Systematic mutational analyses of protein-protein interfaces." Methods Enzymol **202**: 390-411.

Wilson, C. A., J. Kreychman, et al. (2000). "Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores." J Mol Biol **297**(1): 233-49.

Xiong, J. P., T. Stehle, et al. (2001). "Crystal structure of the extracellular segment of integrin alpha Vbeta3." Science **294**(5541): 339-45.

Yang, Z., R. Nielsen, et al. (2000). "Codon-substitution models for heterogeneous selection pressure at amino acid sites." Genetics **155**(1): 431-49.

Young, M. M., N. Tang, et al. (2000). "High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry." Proc Natl Acad Sci U S A **97**(11): 5802-6.

Yue, K. and K. A. Dill (1992). "Inverse protein folding problem: designing polymer sequences." Proc Natl Acad Sci U S A **89**(9): 4163-7.

Yue, K., K. M. Fiebig, et al. (1995). "A test of lattice protein folding algorithms." Proc Natl Acad Sci U S A **92**(1): 325-9.

Zhang, H., K. Huang, et al. (2000). "Crystal structure of YbaK protein from Haemophilus influenzae (HI1434) at 1.8 A resolution: functional implications." Proteins **40**(1): 86-97.

Zuckerkandl, E. and L. Pauling (1965). "Molecules as documents of evolutionary history." J Theor Biol **8**(2): 357-66.

Zwanzig, R., A. Szabo, et al. (1992). "Levinthal's paradox." Proc Natl Acad Sci U S A **89**(1): 20-2.