

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Empirical analysis of sequence alignments as measures of biological conservation

**Permalink**

<https://escholarship.org/uc/item/10j0h3zh>

**Author**

Otillar, Robert Peter

**Publication Date**

2002

Peer reviewed|Thesis/dissertation

**Empirical Analysis of Sequence Alignments as Measures of Biological  
Conservation**

by

**Robert Peter Otillar**

**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: .....

**Copyright 2001**

**by**

**Robert P. Oillar**

## **Dedication**

*In the name of Dorothy J. Otilar, this work is dedicated to the men and women raised during the Great Depression who served our country during World War II. Their hard work created the rich economic and social opportunities that exist for me today.*

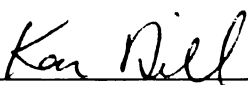
## Acknowledgments

I gratefully thank Drs. C. Anthony Hunt and my thesis mentors Drs. Ken A. Dill, Joe W. Gray, Mark R. Segal, and Manfred D. Zorn for patient guidance and mentoring in completing my PhD in computational biology at UCSF. Drs. Robert J. Fletterick, A. Roger Cooke, and Dennis F. Deen have also shared insightful advice and wisdom at critical junctures, and I offer them my heartfelt appreciation. These senior scientists' attention is in great demand, yet they kindly invested time to pass on scientific knowledge, personal experiences, and professional wisdom to me. They smoothed my path while training me to become an independent scientist. I am forever in their debt. Tireless help and guidance from program director Julie S. Ransom enabled me to succeed in getting my doctorate. My whole family and I recognize and appreciate the help you've given, Julie. I offer my deepest gratitude to Bryan W. Taylor, who contributed countless hours in assistance and discussion on Oracle database construction, data modeling, and administration. Drs. Bertrand R. Huber, Kent E. Duncan, Sarah J. Rice, Herschel V. Wade, Steven. E. Brenner, David J. Storek, and David W. Miller have shared years of advice, support, and commiseration—my grateful thanks to you. The company and support of these dear friends and brilliant young colleagues has been a wellspring of joy and inspired tenacity in my personal and scientific adventures. I also owe a tremendous debt of gratitude and appreciation to *all* the members of my extended family, especially Doris O. Clopper, Donald L. Clopper, Dorothy J. Otilar, Robert P. Otilar Sr., Steven P. Otilar, Laura J. Otilar, Elizabeth A. Otilar, Akitomo Morita, Laura A. Kelly. They have supported and encouraged me throughout the challenge of earning my doctorate and becoming a scientist; without their help and love I probably would not have made it; and I certainly would not have had such a rich life that filled these years with exhilarating experiences.

# **Empirical Analysis of Sequence Alignments as Measures of Biological Conservation**

Detecting signatures of conservation between protein sequences is the primary tool for understanding the biomedical properties and evolutionary relationships of genes and genomes. This work used over 640,000 alignments to empirically analyze two cornerstones of genomic analysis: the pairwise alignment of sequences used to establish putative regions of biological conservation, and percent identity (or similarity) scores used to measure the level of biological conservation present in a pairwise alignment. The focus was on alignments and relationships which cannot be reliably detected with current methods; superfamily versus family alignments were analyzed, as were alignments with marginal significance by expectation score measures, including those in the "twilight zone" below 30% sequence identity. On the biological signal in distant alignments, an extreme value distribution process was found to drive distant homolog alignments, indicating little or no biological information could be gleaned from analyzing those alignments. On percent identity and similarity measures, the length of an alignment was found to strongly predict the number of identical- and positive-scoring residue pairs in the alignment. Commonly used percent identity and percent similarity alignment measures were found misleading or uninformative measures of biological similarity, and the 'HSSP Relationship' between alignment length and identity was reproduced as a mathematical artifact of predicting alignment identity from alignment length. Applications and implications of these results for improving and better interpreting a broad range of methods for calculating sequence alignments between distant homologs were discussed.

Abstract Signature By Thesis Committee Chair

  
\_\_\_\_\_  
Professor Ken A. Dill

## Table of Contents

Copyright.....	ii
Dedication.....	iii
Acknowledgments.....	iv
Abstract.....	v
Table of Contents.....	vi
List of Tables.....	ix
List of Figures and Illustrations.....	x
<b>Chapter I: Introduction.....</b>	<b>1</b>
<i>Motivation.....</i>	<i>2</i>
<i>A Primer on Sequence Alignments and Scores: What They Are, How They Are Calculated.</i> .....	<i>2</i>
<i>The Problems with Distant Homolog Alignments and Alignment Scores.....</i>	<i>8</i>
<i>Our Approach to Solving the Problems.....</i>	<i>11</i>
<i>Our Results for Distant Homolog Alignments.....</i>	<i>13</i>
<i>Our Results for Distant Homolog Scores.....</i>	<i>16</i>
<i>Predictions and a Synthesis of Our Results.....</i>	<i>19</i>
<i>References.....</i>	<i>38</i>
<b>Chapter II: Extreme Values in Distant Homolog Alignments.....</b>	<b>43</b>
<i>Abstract.....</i>	<i>44</i>
<i>Introduction.....</i>	<i>46</i>

1007 1007 1007

<i>Methods</i> .....	49
Sequences, Sequence Databases, and Homology .....	49
Measures and Formulas on Alignments.....	49
<i>Results</i> .....	50
Tabulation of Family, Superfamily Homology Detection .....	50
Comparison of Superfamily, Family Alignment Score Distributions .....	51
Comparison of Superfamily and Non-Homolog Score Distributions.....	51
Single Family Score Distributions.....	52
<i>Discussion</i> .....	53
Superfamily Homologs Appeared Randomly Aligned.....	53
Low-scoring Family Homologs May Be Randomly Aligned.....	54
Sequence Assays Use Non-Biological Alignments for Most Homologs.....	56
Our Results May Apply to Alignment Algorithms in General.....	57
Biological Applications and Better Alignments .....	58
<i>Conclusion</i> .....	60
<i>References</i> .....	67
<b>Chapter III: Empirical Analysis of Percentage Similarity Measures Commonly Used for Understanding Distant Alignments .....</b>	<b>71</b>
<i>Abstract</i> .....	72
<i>Introduction</i> .....	74
<i>Methods</i> .....	78
The Sequence Database .....	78
Calculating and Culling the Alignments.....	78

RECEIVED  
 11/20/11



Measures and Formulas on Alignments.....	79
Analysis Tools and Miscellaneous.....	80
<i>Results</i> .....	80
Weakly Similar Alignment Set Compared to Twilight Zone and Non-Homolog Alignments.....	80
No. of Identities a Linear Function of Alignment Length for Distant Homologs .....	81
No. of Identities a Linear Function of Alignment Length for Non-Homologs.....	81
Alignment Length Predicts Percent Alignment Identity and the HSSP Curve.....	81
<i>Discussion</i> .....	82
Linear Relationship Expected to Hold for SW Algorithms in General .....	82
Basis of the HSSP Relationship.....	84
Percent Identity and Similarity Measures Deprecated.....	84
<i>Conclusion</i> .....	86
<i>Acknowledgements</i> .....	87
<i>References</i> .....	92
<b>Chapter IV: Techniques for Bioinformatics as an Emerging Field .....</b>	<b>96</b>
<i>Entering Bioinformatics</i> .....	97
<i>My Preparation for Research</i> .....	98
<i>What I've found to be Useful</i> .....	102
<i>What I have found Not Useful</i> .....	103
<i>For Building Sustainable Career Skills</i> .....	104
<i>Actions that Prepare for Unanticipated Challenges</i> .....	105
<i>Author's Note</i> .....	106



## List of Figures and Illustrations

Example of basic observable alignment properties .....	26
Raw score distribution for non-homologs, with extreme-value fit .....	28
Raw score distributions for superfamily and family homologs .....	31
Raw score distributions for individual superfamilies and families .....	32
Alignment length vs. identities for distant family and superfamily homologs .....	33
Distribution of distant superfamily and family alignments for common alignment scores .....	34
Percent identity vs. fraction of same-family homologs .....	36
Comparison of family vs. superfamily alignment score distributions .....	63
Extreme value fits to superfamily and non-homolog alignment scores .....	64
Examples of alignment score distributions within individual families .....	65
Extreme value fit to homolog alignment scores .....	66
Alignment length versus number of identities for all PDB90tU alignments .....	88
Alignment length versus percent alignment identity for all non-homolog PDB90tU alignments .....	90

## MOTIVATION

In order to develop new methods for analyzing the Human genome and other sequencing project data, we need to better understand the methods we currently use. Virtually all modern genome project scientists and wet-lab biologists rely on sequence alignments to characterize biological similarities between the sequences they study. As described in this work, alignment methods suggest that most sequence relationships fall in the “distant” homolog<sup>1</sup> category. However, current tools generally fail when attempting to detect or analyze biological similarities between distant homologs. Why? How much biological information do alignments extract from distant homologs? What is the biological meaning of numeric alignment scores for distant homolog alignments? The results described herein help answer these questions<sup>2</sup>.

## A PRIMER ON SEQUENCE ALIGNMENTS AND SCORES: WHAT THEY ARE, HOW THEY ARE CALCULATED

Sequence alignments and alignment scores attempt to characterize *biological* sequence similarities indirectly, by using *residue* similarities (See Figure AP1). A sequence’s residues are the observable outcome of mutations that were accepted or rejected by natural selection during the sequence’s evolutionary history. Residues that govern key biological properties may undergo few changes over billions of years and multiple

---

<sup>1</sup> In other works, the term “homology” is often used as a loosely defined synonym for “similarity”. We do not use that definition. Throughout this work we use the more precise definition that homology means “having a common ancestor”. Thus, “homologs” are sequences that evolved from the same ancestor gene. For further discussion on the use of homology and related terms in sequence analysis see (26).

<sup>2</sup> Unless otherwise noted, our results are based on the PDB90tU alignment dataset that is described in the methods sections of the chapter following this introduction.

speciation events. Others may mutate until they have no detectable similarity to the original ancestor residue. Regardless of retained similarity, two or more *residues* are “homologous residues” when they are descended from the same ancestor residue in the same ancestor sequence<sup>3</sup>. Hence, two sequences are homologous when they contain one or more homologous residues (i.e. if residues in each sequence descended from a common ancestor residue, then some part of the sequences must share a common ancestor). Technically, a residue with little or no change from an ancestor sequence is called “conserved”, and a residue undergoing large changes is called “non-conserved”. Conserved, homologous residues are “mutually conserved”. Usually, the ancestor sequence can’t be directly observed or reconstructed, so one can’t necessarily tell when a residue is conserved. Thus, homologous residues are often called “conserved” when they are similar or identical, regardless of the nature of the ancestor sequence. A primary goal of sequence alignments is to align homologous residues.

By identifying conserved residues, biological similarities between homologs can be discovered<sup>4</sup>. For example, residues that play a key role in determining catalytic activity, regulatory control, or risk for illness can be detected as conserved between sequences from humans, fruit flies, and bacteria. Conservation of a large number of residues, or of key functional residues, suggests conservation of biological properties. Homology between non-conserved residues can often be detected by association with conserved

---

<sup>3</sup> In this work and in the general literature, the term “homolog” refers to homologous *sequences* unless otherwise noted.

<sup>4</sup> Non-homologs may also have residue similarities that correspond to biological similarities. Short signal sequences that target proteins to particular locations in the cell are one example. Non-homologs are commonly treated as lacking biological similarity in sequence-alignment methods-development reports. We also treat non-homologs in this manner unless otherwise noted.

10001120011

homologous residues. Conserved residues are often interspersed with non-conserved residues. A residue sandwiched between residues conserved across several sequences, for example, is probably a non-conserved (i.e. homologous) residue. By mapping homologous residues of all types, alignments delineate entire regions of two sequences that have a common biological origin. Alignments may be inspected to see if particular functional regions have been retained, and alignments may be scored in order to estimate the overall level of conservation.

To generate useful biological information, computational tools need a method to identify homologous residues. For non-homolog sequences, *all* residues paired in *any* possible alignment are, by definition, non-homologous residues. When aligning homologous sequences, many possible alignments would align non-homologous residues. Software tools test many or all of the possible alignments, returning the one with the highest score (the “optimal” alignment)<sup>5</sup>. Thus, the “raw” score used to select optimal alignments provides a bridge between biological similarity and the algorithms used for comparing strings of letters.

Raw alignment scores are models of biological sequence evolution. Raw scores generally add scores from each set of paired residues, plus an insertion/deletion “indel” or “gap” penalty for unpaired residues. The basic model for scoring sequence alignments, introduced by Sellers in 1974, can be thought of as<sup>6</sup>:

---

<sup>5</sup> Theoretically, more than one alignment with the same optimal score may exist. In this case, the choice of which optimal alignment is reported is tool dependent. Further information on this and other aspects of the number of possible alignments can be found in (39).

<sup>6</sup> Sellers’ distance metric (32) is the foundation of sequence distance used in alignment methods, as discussed in (33). This legacy is apparent when inspecting the scoring

*Raw Alignment Score* ~ *Rejected\_Mutations - Indels*

~ *Similarity\_Scores*[aligned residues] + *Gap\_Penalties*

For standard two sequence (“pairwise”) alignments, mutation scores for each two-residue pair in the alignment are taken from a look-up table. Residue similarity score tables are usually calculated from ratios (log-likelihoods) of the mutation and residue frequencies observed in alignments of homolog sequences. Commonly used tables, such as the PAM or BLOSUM matrices<sup>7</sup>, give a positive or negative score to each type of residue mutation. Mutations with a positive score are considered conservative, whereas those with a negative score are considered non-conservative. Indel biology is usually modeled by linear gap penalties: a fixed “open” penalty for starting a gap, plus an “extension” penalty times the length of the gap. Linear penalties are used for reasons of computational efficiency, and are suitable for short gaps<sup>8</sup>. The choice of gap penalty parameters is often heuristic, based on what each user feels gave good results in the past.

A completely “biologically accurate” or “correct” alignment would pair up all homologous residues and no non-homologous residues<sup>9</sup>. Automated alignment

---

methods of local pairwise alignments (see review (3)), multiple sequence alignments (reviewed in (12)), and other methods including profile tools such as PSI-BLAST and IMPALA (31). For broad mathematical treatments of alignment methods and scores, see (39) and (10).

<sup>7</sup> For an informative comparison and discussion of many types of scoring matrices, see (17, 37) and the recent review (14). The biological motivation and statistical justification behind the original PAM and now-standard BLOSUM matrix series are found in (7) and (15).

<sup>8</sup> For an investigation of gap penalties versus biological reality, see (5) and a more statistical treatment in (2).

<sup>9</sup> The study of accuracy in alignments has gained increasing attention in recent years; for an introduction to data-driven investigations of sequence alignment accuracy versus structural alignments, see ((8, 30, 34)) and references therein. See also work and discussion on near-optimal alignments (e.g. (10, 16, 22, 36, 38, 39)).

techniques can have difficulty distinguishing conserved residues from high-scoring, non-homologous residues. Thus, when homologs are aligned some residues may be paired that do not reflect biological conservation or residue homology. In practice, as alignment scores decrease, so does overall alignment accuracy. Varying descriptions of alignment “error”, “inaccuracy”, or “incorrectness” are occasionally used to estimate how well the residues paired by alignment agree with those of an accurate alignment. Much more commonly, a general-purpose alignment score is used to estimate the degree of similarity between two sequences.

For clarity, we emphasize that studies of homolog alignments are about *alignments as computed by current methods*, rather than *the existence or information contained in a correct alignment*. Our work, for example, addresses the ability of current methods to calculate alignments and scores that reflect biological similarity. This is a necessary limitation. If two sequences are homologous, then they have homologous residues that can be paired in a correct alignment. However, we cannot go back in time and learn the true evolutionary history of each residue in a naturally occurring sequence. Current evidence may indicate that two residues are similar by various criteria, but that does not prove they are homologous. An alignment of similar residues is simply the “best guess” available to us. Improving current methods enables “better guesses.” Studies of alignments, such as ours, thus deal with alignments as computed by current tools and the scores used to extract biological information from those computed alignments.

Extreme value scores are a primary tool for distinguishing homologs from non-homologs. Extreme value statistics were developed by Gumbel in the 1950’s for estimating the



probability that at least one flood in a given year would be a severe flood<sup>10</sup>. Unlike the normal distribution, which estimates probabilities for a *sum* of several random numbers, the extreme value distribution estimates probabilities for the *maximum* of several random numbers. For two non-homologs, the score of each possible alignment is a random number, and the reported alignment's score is the maximum of those numbers. (This highest-scoring possible alignment, rather than the most accurate one, is the "optimal" alignment that methods like the Smith-Waterman local alignment algorithm report.) Therefore, alignment scores for non-homologs follow an extreme value distribution (as shown in Figure RNH1). Extreme value statistics estimate the probability that two non-homolog sequences would generate a raw score equal-or-higher than the score of an alignment of interest. If the probability is very low, then we reject the hypothesis that non-homologs generated the alignment of interest. If two sequences are not non-homologs, we conclude they are homologs. Hence, rejection of the preceding hypothesis means we have decided that the alignment score is "significant" and that homology between those sequences has been "detected". Longer sequences are more likely to have a high alignment score by chance, as are sequences with unusual frequencies of rare (or common) residues. Hence, modern extreme value statistics enable inclusion of corrections for sources of bias such as sequence length and residue frequencies. When a search sequence is compared to a sequence database, the "expect" or "e-value" score reported for an alignment equals the number of database sequences ("hits") expected to have an equal-or-higher score than that alignment *without being homologs of our search sequence*. These scores have been empirically verified as accurate or conservative.

---

<sup>10</sup> For details (13); for a brief mathematical synopsis see also(11).

Hence, given an alignment with an e-value of 100, we would expect to see 100 non-homolog hits having raw scores greater-or-equal than the raw score of that alignment. A “expect” or “e-value” score of  $10^{-2}$  basically means that there is a 1% chance that an observed alignment is due to non-homolog sequences.

Percent identity scores are used more often than statistical scores when measuring similarity between two sequences<sup>11</sup>. Percentage scores divide a basic mutation score (without gap penalties) by a measure of length (See Table PSF1). These scores do not involve nonlinear calculations or gap penalty selection, and admit few or no choices about the mutation score parameters. Percent identity scores are widely held as intuitive and easy to understand<sup>12</sup>.

#### **THE PROBLEMS WITH DISTANT HOMOLOG ALIGNMENTS AND ALIGNMENT SCORES**

Many biological relationships between sequences are distant and hard to characterize with alignments. Almost by definition, distant homologs lack residue conservation that is easily characterized by any current software tools. Conceptually, homologs are “distant” when they are separated by great evolutionary distances. In practice, homologs are considered “distant” when they cannot be distinguished from non-homologs on the basis of their alignment scores. Sequence alignment scores often define distant homology: homologs below a alignment score cutoff are considered distant, those above the cutoff are not. Common cutoffs include a 30% sequence identity “twilight zone” cutoff and e-value cutoffs of  $10^{-2}$  to  $10^{-5}$ .

---

<sup>11</sup> Percent identity scores are vociferously deprecated by some statistical bioinformaticians. Nonetheless, percentage scores remain frequently used for judging sequence similarity, as noted by the comparison of percentage and e-value scores in (41).

<sup>12</sup> For a straightforward interpretation of percent identity scores as exponentially related to mutational distance, see the classic book (9).

*Most* homologs are distant homologs. In other words, given any two sequences, we generally cannot tell if they are homologs<sup>13</sup>. Homology is more difficult to detect between sequences that are more evolutionarily distant. Many sequences are clustered into “families” of homologous sequences that also share similar biological characteristics; when sequences in different families are homologous, those families are clustered into “superfamilies”<sup>14</sup>. Homologs in different families, “superfamily-level homologs”<sup>15</sup>, are generally believed to be more evolutionarily distant than homologs in the same family. Recent studies indicate that only 4% of superfamily-level homology relationships are detected by alignment methods<sup>16</sup>.

Improved statistics and multiple-sequence methods have generally increased our ability to reliably detect the homologies that were already evident by other means. When two sequences are homologous to a third, alignments containing all three often make it possible to determine if the first two are also homologs. (Technically, one homologous

---

<sup>13</sup> Numerous studies make this point; see (6, 18, 21, 24) and references therein. Techniques to increase the effectiveness of alignment searches for particular applications are discussed in (25). Using standard pairwise methods and default scoring parameters, our analyses confirmed over 50% of homologs as distant under e-value ( $10^{-2}$  and  $10^{-5}$ ) and twilight zone (30% percent alignment identity) cutoffs.

<sup>14</sup> For the SCOP database used in our work, homologs are classified into the same family based on percent sequence identity or strong functional and structural similarity (19). Hence, the family classification provides a somewhat fuzzy mix of sequence and non-sequence similarity information. Superfamily classification is more rigorous: if two sequences are homologous by any criterion, they are in the same superfamily. Because SCOP sequences each consist of only a single protein domain, homology is transitive for SCOP sequences. Thus, each sequence is classified in exactly one family contained in exactly one superfamily.

<sup>15</sup> Confusingly, some authors use “superfamily homologs” to indicate only homologs in different families, while for others all family homologs are also superfamily homologs. Throughout this text, we use the former definition.

<sup>16</sup> Detection of 4% or less on the SCOP reference database is shown for pairwise, profile, and hidden Markov model methods in (18). Our data confirmed these results for pairwise methods.

residue shared by all three sequences implies they are homologs.) When transitivity is used, the final question, “is sequence A homologous to sequence B,” is answered similarly well by multiple searches with traditional pairwise techniques, modern pairwise alignments, or multiple-sequence methods<sup>17</sup>.

Even when we know two sequences are distant homologs, it is difficult to calculate alignments or scores that reflect the known biological similarity. Recent systematic studies have described the performance of widely used alignment tools on distant homologs<sup>18</sup>. These studies confirm that alignment accuracy drops off sharply as alignment scores approach cutoffs for distant homology. A similar trend is observed for the correlation between alignment scores and measures of biological similarity. Detailed plots in these studies compare x-axis measures of alignment-based sequence similarity, including percent alignment identity and e-values, to y-axis measures of biological similarity, including agreement with structure-based alignments, three-dimensional RMSD structural similarity scores, beta-sheet and alpha-helical secondary structure similarity, and functional similarity based on enzyme classifications and protein function. A decrease in agreement between alignment accuracy, alignment scores, and various measures of biological similarity is consistently observed for distant homologs. Some tools are found to perform better than others on particular types of homolog data, but all methods perform increasingly poorly as alignment scores drop. For all methods tested, on

---

<sup>17</sup> Performance benchmarks and analyses have become a popular topic in recent publications. As examples, see (18), and references therein, for comparisons of pairwise, iterative-pairwise, profile, and hidden Markov methods. (6) set the standard for modern performance benchmarking, and illustrates the near equivalence of traditional Smith-Waterman and modern pairwise methods.

<sup>18</sup> See, for example, comparisons of sequence- and structure- alignments in (30, 34, 40) and references therein.

average over 50% of the residues aligned for the lower-scoring distant homologs were inaccurately aligned. That is, most of the aligned residues were actually not homologous residues. The reasons for the breakdown in alignment accuracy remain largely unexplained.

As these difficulties suggest, advances in measuring similarities between distant homologs are needed. The challenge is to identify the types of changes that should be made to improve current alignment methods and scores for use with distant homologs.

### **OUR APPROACH TO SOLVING THE PROBLEMS**

I believe that the universal breakdown of alignment-based methods on distant homologs suggests a common point of failure. Some assumption in how new alignments are calculated, or how fixed alignments are scored, seems at fault in all methods. One way to detect the cause of this failure is to better understand how representative methods agree, or disagree, with biological reality. As long as some aspect of biology continues to influence the alignment process and the resulting scores, biological information remains to be extracted with those methods. Alignment methods that reliably and correctly pair homologous residues will enable new biological insights. Scoring methods that reliably quantify an aspect of biological similarity that is present in an alignment would be immediately useable. Hence, two objectives that enable improvements are: identify when representative tools extract biological information from distant homolog alignments; and clarify the physical aspects of alignments reflected by current similarity scores.

Towards these two objectives, we have studied pairwise alignments and their scores. Pairwise alignment tools provide the most direct route to understanding the distant homolog problem. Virtually all modern alignment methods use alignment and scoring

concepts co-opted from the pairwise comparison literature. Many multi-sequence methods calculate actual pairwise alignments as intermediate steps. Variants of pairwise scores, such as percent identities and e-values, are used with alignments from all methods. As the simplest tools for computing sequence similarity, pairwise alignment methods may isolate the essential elements of sequence comparison that fail for distant homologs.

We have taken a data-driven approach to studying pairwise alignments and scores. We aligned real biological sequences to establish large-scale data sets. We compared sequences that share known biological similarity—homology—but uncertain sequence similarity. Consistent with other recent analyses of alignment tools, we performed all-against-all alignments on a standardized reference-sequence database. Unlike previous studies, our general exploration of the data required the flexibility to isolate and analyze many distinct aspects of alignment similarity. Exploring data in this way is difficult because all-against-all alignment generates millions of alignments, and tens of millions of data points. Relying on available resources to write software to parse and analyze aspects of the data for each exploratory question seemed infeasible. Historically, limitations in computational resources and training have limited large-scale, data-driven “big data” research to asking narrow questions with custom-written software. Fortunately, advances in computing technology are beginning to make flexible, big-data research accessible to computational biologists with expertise in data-modeling, statistics, and database management systems. In our work, data was warehoused in a commercial relational database and analyzed using structured query language tools and standard statistical suites. This type of data-driven analysis provides a new, promising methodology for

biological research. Moreover, the techniques we developed allowed us to identify properties of pairwise alignments that remained unreported by other scientists during three decades of previous research on alignments.

#### **OUR RESULTS FOR DISTANT HOMOLOG ALIGNMENTS**

Virtually all alignments computed for superfamily homologs appeared to reflect no information about biological similarity, whatsoever. Overall, the information captured by distant homologs' alignments appeared to depend on their evolutionary distance, not just their score. Judging the biological information in distant homolog alignments—without knowing the accurate alignment—appeared to open the door to improved distant-homolog methods.

In our analysis, we identified extreme-value distributions in superfamily homolog alignments (see Figure RSF1, graph A). Extreme-value behavior suggests that these homologs followed the same alignment process as non-homolog sequences. Extreme value distributions are commonly observed for non-homolog sequences as well as for randomly generated sequences. When an extreme-value process is observed, the implication is that the alignment is “random”. By random we mean: “shuffling the residue-types for each sequence would not be expected to change the number of correctly aligned residues”. Regions aligned, residues paired, and scores calculated for random alignments can reflect biological similarities forced into the calculations by similar sequence lengths or unusual residue frequencies. However, random alignments themselves do not add biological information beyond what was available without aligning the sequences. For randomly aligned sequences, some homologous residues may be paired by chance. Overall, however, such alignments are expected to be almost

entirely inaccurate. Randomly aligned sequences have “random accuracy”: mutating the sequences’ residues many times without regard to conserving biological similarity, and then re-aligning them, would not be expected to reduce the number of correctly paired residues. Thus, “an alignment added no information” is the only new thing that a random alignment can tell us about the biological similarities of two homologs.

Extreme values in distant homologs are not previously reported, but they are unsurprising in hindsight. Our observations are consistent with the idea that homologs with few conserved residues will have fewer high-scoring residues in a correct alignment. If the residues paired in a random alignment give a higher similarity score than those in all non-random alignments, then it will be the random alignment that is reported by an alignment program. Low scoring alignments are expected for many distant homologs. That expectation is consistent with observing random alignments and an extreme value score distribution for those homologs.

The information captured by distant homolog alignments may depend on evolutionary distance, not just alignment scores. Distant homologs were separately analyzed as family- and superfamily-level homologs (the former being more closely related than the latter). Low-scoring family-level homologs’ alignments gave different results than those of similar-scoring superfamily-level homologs. Virtually all superfamily homologs were low-scoring. These very distant homologs followed an extreme-value distribution, suggesting they are randomly aligned (see Figure RNH1). However, the situation was less clear for family-level homologs. Family-level homologs did show a peak at low scores. That peak that may indicate an extreme-value distribution mixed with other distributions or alignment processes. However, when inspected individually, only some



families exhibit a low-score peak (see Figure RSF1). This variation was not observed when individual superfamilies were inspected. The lack of a peak suggests that some families' low-scoring alignments do not reflect an extreme value process. A non-extreme-value process suggests these are not random alignments; biology is influencing some family alignments, even at scores where all superfamily alignments appeared to be random.

Random homolog alignments may enable new insights into how to measure distant homology. Alignment method analyses typically compare homolog alignments, which are assumed to measure biological similarity, versus non-homolog alignments, which are assumed not to reflect biological similarity. As noted, superfamily alignments appeared to be randomly aligned, while many family alignments did not. To increase accuracy, we suggest that analyses relying on current alignment tools should remove superfamily homologs from alignment data assumed to reflect biological similarity. For studies attempting to explore how alignment algorithms fail to capture biological similarity, alignments that failed to capture biological similarity (randomly homolog alignments) may provide better controls than alignments lacking any biological similarity to capture (non-homolog alignments). For these studies, family and superfamily alignments should be compared. Furthermore, comparisons between family- and superfamily- homologs may use subsets selected for similarly computed similarity scores *and* similar biological properties, like fold or catalytic activity. Through this process, confounding factors can be minimized or eliminated, because both subsets are known *a priori* to share particular biological similarities. These types of comparisons may also be useful between roughly

defined extreme-value and non-extreme value family alignments, though such subsets may be challenging to isolate.

Extreme-value fits to homolog data may provide a positive assay for capturing more biological information with improved tools. By measuring a tools' ability to decrease the extreme-value behavior of known distant homologs, improvements in existing tools can be guided and validated. For example, if we take two methods and plot their distant homologs' alignment scores, we can fit an extreme value distribution to those scores. We can then compare the area under the two extreme-value curves<sup>19</sup>. A decrease in area indicates that more biology was captured by an alignment method. The method with smaller area (fewer alignments) following an extreme value distribution is expected to have fewer random alignments. So, comparing methods by their homolog extreme value distributions provides a test for when we are capturing more, or less, biological information in the alignments. Unlike existing assays that require sequences of known structure, extreme-value homolog tests can be applied to sequences where their structure is unknown. Hence, extreme-value homolog tests should enable alignment methods for characterizing homologs without known structures, such as G-protein coupled receptors and other transmembrane or non-globular proteins.

## **OUR RESULTS FOR DISTANT HOMOLOG SCORES**

Percent identity scores took on new and unexpected meanings when applied to distant homologs. Our data indicated that, *for distant homologs*,

---

<sup>19</sup> Non-homologs provide a control against trivial false-positives. Non-homologs should generate similar extreme-value distributions under both algorithms.

- Percent alignment identity is a random number that is not an indicator of similarity in any usual sense;
- Percent sequence identity scores simply tell if an alignment is full length.

Our results on percent identity measures changed how we interpret other historical works.

For distant homologs, the number of identities and the alignment length are highly correlated (see Figure AID1). For both family and superfamily homologs, identities were well described as a linear function of alignment length:

$$\text{Identities} = m * \text{AlignmentLength} + b \pm \text{error}$$

Fitting this equation to distant homolog alignment data, we found that alignment length explained 90% of the variance in identities (fit parameters were  $m = 0.25$  identities per residue,  $b = 2.9$  residues,  $\text{error} \sim 0 \pm 2.4$  residues)<sup>20</sup>.

Effectively, these two alignment properties reflect a single interchangeable number for distant homologs. In light of this equivalence, percent alignment identity can be viewed as:

*Percent Alignment Identity =*

$$\frac{\text{Identities}}{\text{AlignmentLength}} \cong \frac{\text{PredictedIdentities}}{\text{AlignmentLength}} \cong 25\% + \frac{2.9 \pm \text{error}}{\text{AlignmentLength}}$$

Hence, all distant homologs with long alignments give approximately 25% alignment identity. Shorter alignments deviate from the baseline when the difference in a few

---

<sup>20</sup> One might expect a fit to yield zero identities at zero alignment length. However, current tools always report positive or identical residues at both alignment edges. (Score optimization means negative scoring edge-residues would be trimmed before an alignment is reported). Because they can align at either side of the gap, residues near gaps are also biased to be positive or identical. These biases contribute to a non-zero intercept.

residues (2.9±2.4 residues) becomes noticeable when dividing by a short alignment length. Hence, percent alignment identity appeared irrelevant as a measure of biological similarity for distant homologs. Said differently, if pairwise alignments are calculated for a sequence with two of its distant homologs, the first showing 18% alignment identity and the second showing 28%, there is no reason to believe the second homolog is more closely related to the sequence than the first homolog.

Percent sequence identity for distant homologs can be cast as:

*Percent Sequence Identity* =

$$\frac{\text{Identities}}{\text{MaxAlignmentLength}} \cong \frac{\text{PredictedIdentities}}{\text{MaxAlignmentLength}} \cong 0.25 \times \frac{\text{AlignmentLength}}{\text{MaxAlignmentLength}}$$

where *MaxAlignmentLength* is the length of the shortest sequence to be aligned, i.e. the maximum possible alignment length. Hence, percent sequence identity appeared to have a simple physical interpretation for distant homolog alignments: the percent aligned, divided by four (up to a maximum of 100%).

In light of our results, it becomes clear that the HSSP curve<sup>21</sup> is not related to structural or biological similarity between sequences. The HSSP curve is essentially the upper boundary of alignment data from a scatterplot of alignment length versus percent alignment identity (See Chapter III, Figure 2). This curve was introduced to describe a boundary-relationship between sequence and structural similarity, and has been cited and studied in several previous works. Given identities' dependence on alignment length, the

---

<sup>21</sup> The HSSP curve is noted as a “principal result” of the classical work (28). As an example that the HSSP curve continues to attract scientific attention, we simply note that his been variously criticized, reproduced, updated, or otherwise discussed in references including (1, 6, 21, 27).

HSSP curve is equivalent to plotting “X” vs. “1/X” with an error term. Hence, the HSSP curve appeared to be a mathematical artifact of identities’ dependence on alignment length. We concluded the HSSP curve does not merit further study or use.

## **PREDICTIONS AND A SYNTHESIS OF OUR RESULTS**

We have discovered and explained several modes of failure for alignment methods and scores on distant homologs. Ultimately, the hindsight of success in characterizing distant homologs will be required to prove why current methods are failing. However, using insights gained during our analyses, we advance several points and hypotheses.

Distant homolog alignments appear to fall into two basic types:

I. Alignments with a high-accuracy core bordered by long, inaccurate extensions

II. Random alignments

Type I (“anchor”) alignments encompass the closer distant homologs; these alignments extract new information about biological conservation from the homologs’ sequences. Type II (“random”) alignments encompass the more remote distant homologs; these alignments are almost entirely inaccurate and extract no new information about biological conservation from the homologs’ sequences. We now discuss each type in more detail.

Random alignments explain observations about distant homologs that appeared to follow the same alignment process as non-homolog sequences. Like non-homolog alignments, the raw scores of random alignments follow an extreme-value distribution. Almost all superfamily homologs, and some low-scoring family homologs, produced characteristic signatures suggesting their alignments were random. As previously discussed, the residues in random alignments are likely to be paired without regard to which residues

were homologous. These alignments are almost wholly inaccurate and their scores do not measure the aligned homologs' biological similarity.

Anchor alignments explain observations about family-level distant homologs that did not appear randomly aligned. Most of the residues aligned for same-family distant homologs appear incorrectly but perhaps not randomly aligned. In our hands, the identity-length relationship for distant family homologs suggested that their alignments extended by adding one identical residue for each four residues that were added. Non-homolog alignments behaved similarly, suggesting that the identity-length relationship was due to non-biological alignment extensions. However, many distant family alignments did not appear to follow an extreme-value process that would indicate they were randomly aligned. These observations are consistent with a high-accuracy core alignment that anchors a surrounding inaccurate alignment. Such an alignment would not be simply the maximum of many possible random alignments, and hence would not be expected to follow the same extreme-value score distribution as randomly aligned distant homologs. Most of the length of the alignment, however, would be in the inaccurate extensions to the anchor. By extending an average of four residues for every identical residue aligned, long inaccurate extensions would dominate the identity-length relationship in the overall alignment. In some cases, these inaccurate regions may not be physically contiguous with the anchor; gaps may intervene. The final alignment could even be a tiling of accurately and inaccurately aligned segments that were independently calculated. In all cases, an accurate anchor would greatly constrain the number and type of alignments that are possible for other regions of the sequences. Biologically accurate core alignments with large inaccurate extensions have been previously observed in distant homologs, and are

consistent with alignment edge-wander theory<sup>22</sup>. The anchor hypothesis advanced here predicts that the inaccurate extensions will be statistically distinguishable from the conserved anchor. Unlike the anchor, the inaccurate extensions should maintain an approximately fixed ratio of aligned residues to identical residues. (A 4:1 ratio would be expected under the conditions used in our alignment calculations). The anchor, however, should contain accurately aligned, conserved residues that would be detectable as unusual densities of high-scoring residue pairs. These high scoring residues may be distributed throughout the anchor or clustered at the ends. Sliding windows, as used for hydrophobicity plots, or formal scan statistics may be used to test these predictions.

Distant homolog alignments are probably not accurate enough to interpret the biological meaning of their alignment scores. Most residue pairs in most distant homolog alignments are incorrectly aligned. Moreover, these incorrectly aligned residues are not paired based on biological similarities such as convergent evolution. The biological roles of these residues are not a factor in how they are aligned. Hence, most residue pairs in most distant homolog alignments contribute only random noise to alignment scores. For random alignments, all residues contribute pure noise, and the resulting alignment scores are random. For anchor alignments, residue-pair scores from an accurately aligned region may be biologically meaningful. Standard alignment scores for distant homologs' anchor alignments will include a large random-noise error term, as most residue pairs are expected to be in the incorrectly aligned extensions. Some scores may accurately measure alignment properties. Percent sequence identity, for example, measures how much of the

---

<sup>22</sup> Alignment errors near gaps are generally expected, particularly for distantly related sequences. For recent results on this topic, see (30) for distant homologs and (29) for close homologs. Edge-wander theory is treated in (16).

maximal alignment length was achieved. However, biological information that is lost in the alignment process cannot be regained in the scoring process. Alignments' divergence from biological reality appears to prevent alignment scores from providing a reliable measure of biological similarity.

The distinction between random alignments and anchor alignments makes testable predictions about studies of distant homologs. These predictions are based on changes in the level of signal and noise in distant homologs' alignment scores. As one prediction, trends in agreement between biological similarities and alignment scores should be different for distant family and superfamily homologs. These differences should reflect the changes in noise versus score that occur in random and anchor alignments. Family homologs with lower alignment scores may coincide with shorter anchor alignments and lower accuracy alignments. Hence, the absolute level of biological information in family homolog scores may decrease as the scores themselves decrease (i.e. lower scores may indicate lower information, rather than lower biological similarity). Anchor alignments' inaccurate extensions will also contribute noise in their alignment scores; the extensions' contribution may increase at lower values for some alignment scores. The average level of biological information, and noise, in distant superfamily homologs should both be constant at different alignment scores. Virtually all superfamily homologs should generate random, pure-noise alignments and scores. Hence, the agreement of biological similarity and alignments (or their scores) should not strongly depend on alignment score for distant superfamily homologs. In contrast, this agreement is predicted depend on the magnitude of the alignment score for family homologs. A second prediction is that trends-lines in population studies should demonstrate a qualitative change at scores where



the proportion of superfamily homologs increases. Alignment scores are often compared to other biological properties in population studies that include both distant family and distant superfamily homologs. Such “mixed population” studies analyze the signals of random and anchor alignments into a single trend<sup>23</sup>. As alignment scores decrease, mixed population studies of distant homologs generally observe growth in the discrepancy between most alignment scores and most other measures of biological similarity. Noise from inaccurately aligned regions in anchor and random alignments should both contribute to these discrepancies. However, the contributions of random versus anchor alignments should be distinguishable. The percentage of distant homologs that are superfamily, rather than family, homologs increases rapidly as most alignment scores decrease (see Figures ISF1 and PFH1). Our results predict that an increase in the percentage of distant homologs that are superfamily homologs, and therefore random homologs, should produce a corresponding drop in the observed agreement between alignment scores and biological similarity scores. The solid curves in Figure ISF1 represent the fraction of homologs expected to contribute random alignments and scores that are effectively pure noise. The dashed curves illustrate the decrease in homolog alignments that contain biological information in mixed population studies. Virtually all signal of biological similarity must come from non-random (i.e. family) alignments. The percentage of alignments that are from family-level homologs (dashed curves) provides an upper bound to the fraction of alignments that may be anchor alignments. (Some family alignments may be random alignments). These anchor alignments and their scores may contribute biological information. The dashed curves thus provide a practical upper

---

<sup>23</sup> As examples, see the systematic comparisons of alignment scores and biological similarity in (30, 40, 41) and references therein.

bound to the fraction of distant homolog alignments that may contribute biological information in mixed population studies. Thus, distant-homolog mixed-population studies combine trends from two processes<sup>24</sup>: first, the changing fraction of alignments that contribute pure noise (random alignments, solid curves) versus alignments that may contain biological signal (anchor alignments, dashed curves); second, the change in the information contained in non-random alignments at lower sequence similarity. The distinction between these two processes is not generally made in studies of distant homolog alignments. We suggest that the first process may dominate the disagreements between alignment scores and biological similarity reported to occur in mixed population studies of distant homolog alignments. Other differences in when signal, or noise, occur in family and superfamily alignments also may have distinguishable statistical characteristics. Overall, we suggest that alignment errors are probably the primary cause of distant homologs' sudden drop-off in agreement between alignment scores and biological similarity scores.

The difficulties in computing similarities for distant homologs appear to arise because current scoring methods don't adequately model the biological events that occur at large evolutionary distances. This can be changed. Moving forward, we suggest that improvements in alignment raw scores should be the primary focus in developing methods to characterize distant homologs with alignment methods. Advances in methods that rely on accurate pairwise alignments, such as multiple-sequence methods and

---

<sup>24</sup> Statistical analysis of error trends by mixture-distribution fitting is beyond the scope of this work. However, we refer the interested reader to references on mixture-models and generalized linear models (20, 35) with an initial basis for a two-component model:

Population 1: *Biological Similarity Score*  $\sim$  *Alignment Score* \* (1 + *Error*<sub>1</sub>)

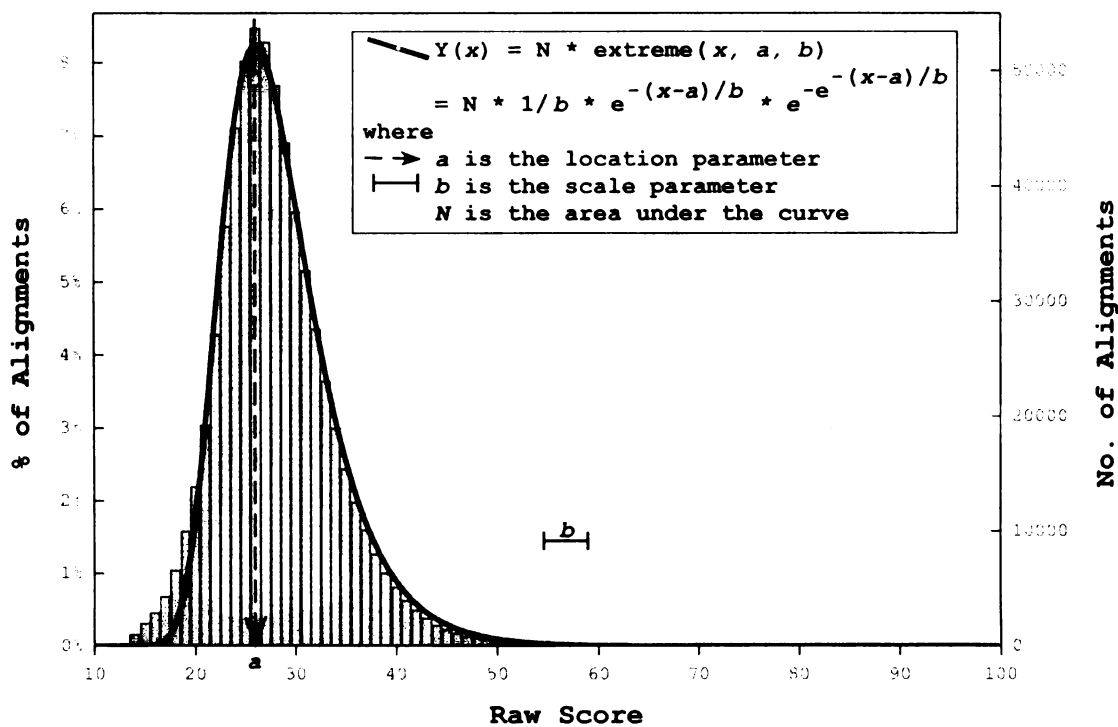
Population 2: *Biological Similarity Score*  $\sim$  *Error*<sub>2</sub>

improved alignment significance statistics, appear less important. As an immediate next step, distant family homolog alignments should be studied to clarify sequence characteristics that signify a boundary between accurately and inaccurately aligned regions. These characteristics may enable improved gap penalties or raw alignment scores that provide higher accuracy alignments for distant family homologs. Removing accurately aligned regions and re-aligning the remainder may allow additional accurately aligned regions to be reported and identified. Higher accuracy alignments would enable functional characterization of distant homologs, by removing uncertainty in whether aligned residues are truly conserved. Identifying an accurately aligned region proves homology; new measures for identifying accurately aligned regions may also improve homology detection. Once tools are developed for identifying regions that are accurately aligned by current alignment scores, it may become feasible to reliably detect and characterize the biological similarities of distant family and superfamily homologs.



single *Gap Penalty* is in angle brackets, calculated using BLAST defaults: standard linear gap penalties, -11 for opening a gap plus -1 for each one-residue extension of the gap. C-E. Breakdown for calculating the *Identities* (13), *Positives* (22), and *Alignment Length* (50) for the alignment in A. We note that the Query and Sbjct sequences aligned are respectively: formate dehydrogenase, 187 residue sequence length, (SCOP domain ID d2nac1); and D-glycerate dehydrogenase, 130 residue sequence length, (SCOP domain ID d2nac1). These sequences are homologs from same SCOP family: formate/glycerate dehydrogenases, NAD-domain.

**Raw Score Distribution for Non-Homolog Alignments, with Extreme-Value Fit**



**Figure RNH1.** Raw score distribution for non-homologs, with extreme-value fit.

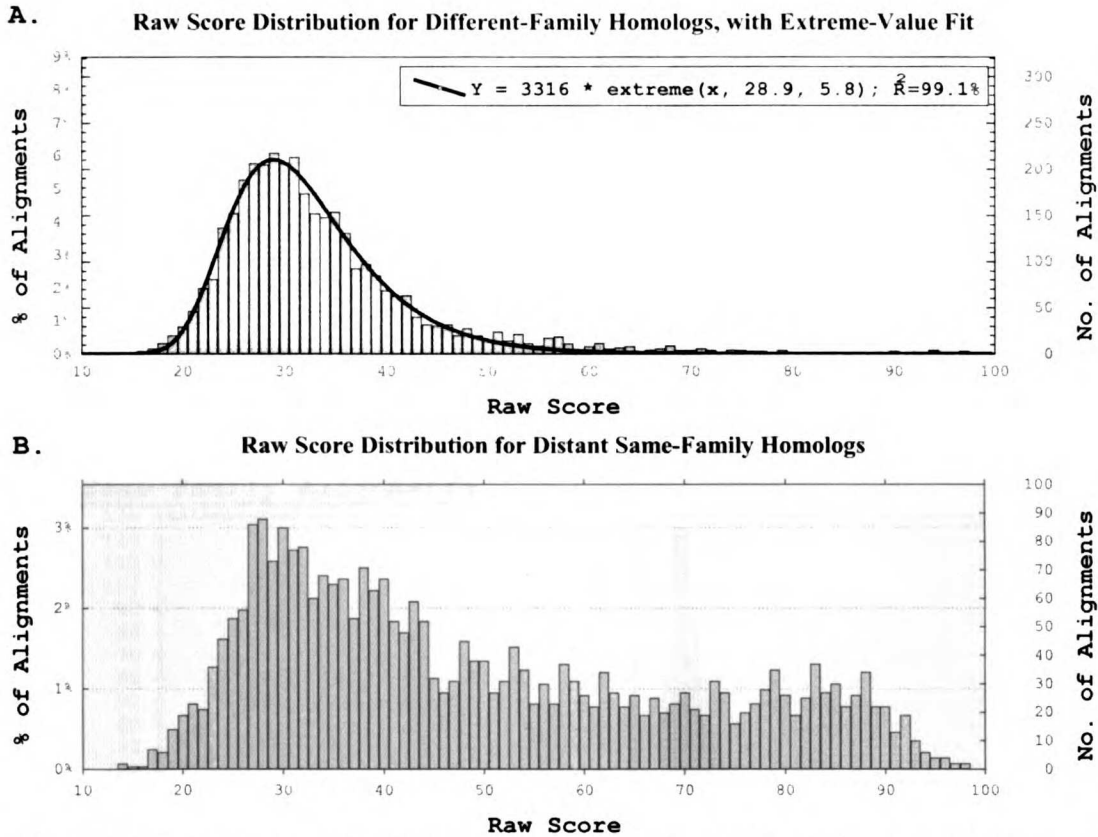
Histograms and extreme value distribution curve fit for raw alignment scores for all non-homolog alignments from the PDB90tU dataset. Extreme value fit data was taken as raw scores with histogram bin counts from this graph (bin width of one point raw score). Fit was performed by nonlinear estimation on  $\{Y = N * \text{extreme}(X, \text{location}, \text{scale})\}$  using quasi-Newton fitting. The curve fit yielded  $Y = 624971 * \text{extreme}(X, 26.0, 4.4)$ ;  $R^2=99.4\%$ . For the fit as shown,  $X = \text{Raw Score}$ ,  $Y = \text{No. of Alignments}$ . For database searching and p- or e-value estimation, the  $X$  term for the extreme value fit generally a modified raw score that adds a correction factor for the length of the two sequences aligned. We focused on analyzing alignment properties, rather than detecting homologs. Hence, we used raw scores as directly observed from the alignments to allow direct

attribution of cause and effect in our analysis. (In our hands, raw scores explained approximately 98% of the variance in  $\log(e\text{-value})$  for distant homologs. Adding standard correction factors to the raw score was not expected to change our results.) See (4) and references therein for raw score correction factors in extreme value statistics. See (3) for an introduction to approximations of the extreme value distribution often used in database searches.

<u>SCORE NAME</u>	<u>SCORE FORMULA</u>
<i>Percent Sequence Identity</i>	$= \frac{Identities}{MaxAlignmentLength}$
<i>Percent Alignment Identity</i>	$= \frac{Identities}{AlignmentLength}$
<i>Percent Sequence Similarity</i>	$= \frac{Positives}{MaxAlignmentLength}$
<i>Percent Alignment Similarity</i>	$= \frac{Positives}{MaxAlignmentLength}$
<i>Percent Aligned</i>	$= \frac{AlignmentLength}{MaxAlignmentLength}$

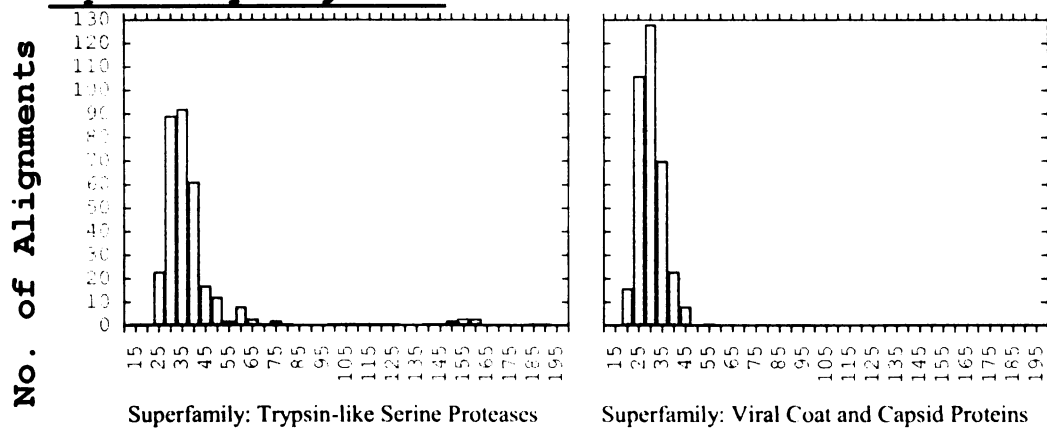
**Table PSF1.** Percentage score formulas. Alignment percentage scores are based on a ratio dividing an observable alignment property by a measure of length. *AlignmentLength* is the number of residue positions aligned in an alignment (residues aligned with gaps are not counted). *MaxAlignmentLength* is the longest possible length of the alignment, equal to the number of residues in the shortest sequence given to the alignment algorithm. *Identities* is the number of alignment positions where all residues are identical. *Positives* is the number of alignment positions with a positive similarity score, where positive is determined by the residue similarity scores used to calculate the alignment.



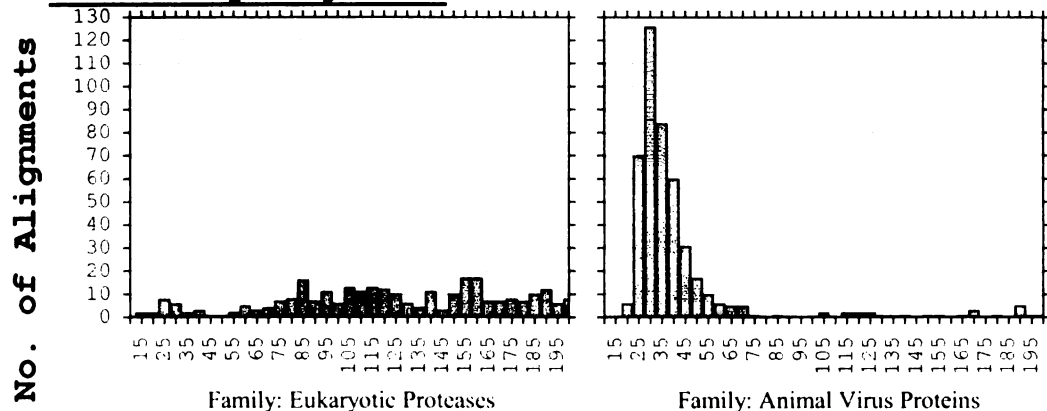


**Figure RSF1.** Raw score distributions for superfamily and family homologs. Histograms of raw alignment scores for alignments from the PDB90tU dataset. **A.** All superfamily homolog alignments (i.e. homologs not in the same family) with extreme value distribution curve fit. Fit parameters and variance explained by the fit ( $R^2$ ) are shown in the graph inset. Extreme value fit data was taken as raw scores with histogram bin counts from this graph (bin width of one point raw score). Fit was performed by nonlinear estimation on  $\{Y = N * \text{extreme}(X, \text{location}, \text{scale})\}$  using quasi-Newton fitting. **B.** All same-family homologs below a distant-homolog cutoff at BLAST e-value  $\leq 10^{-5}$ .

**A. Superfamily Alignments**



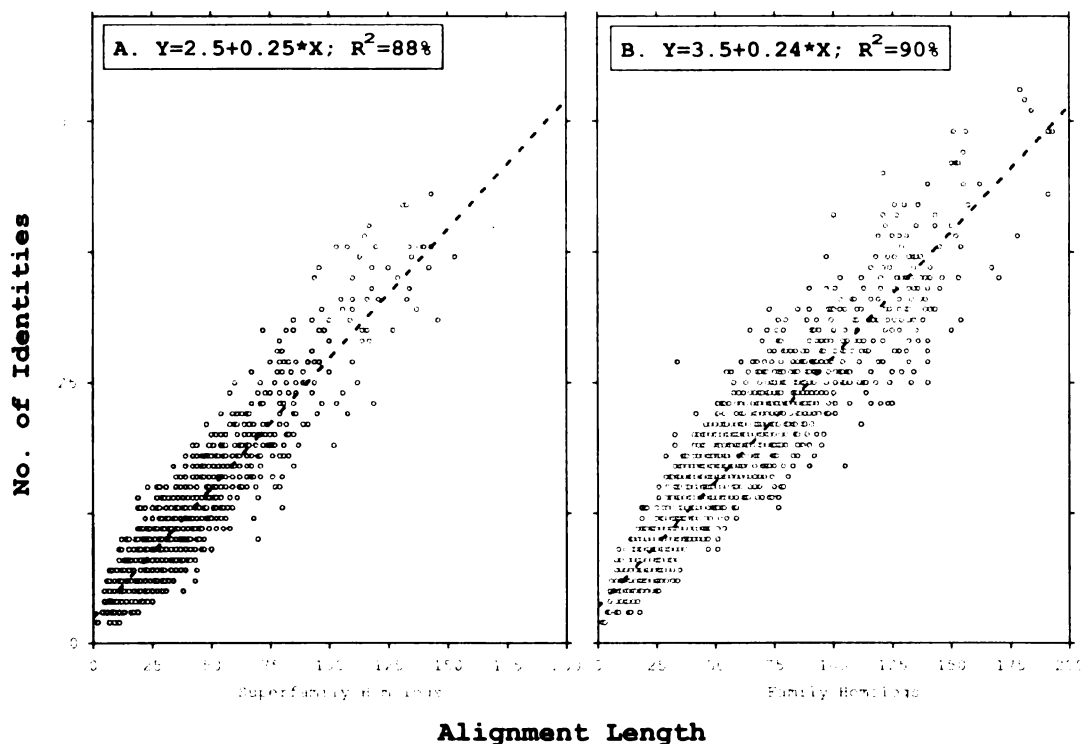
**B. Same-Family Alignments**



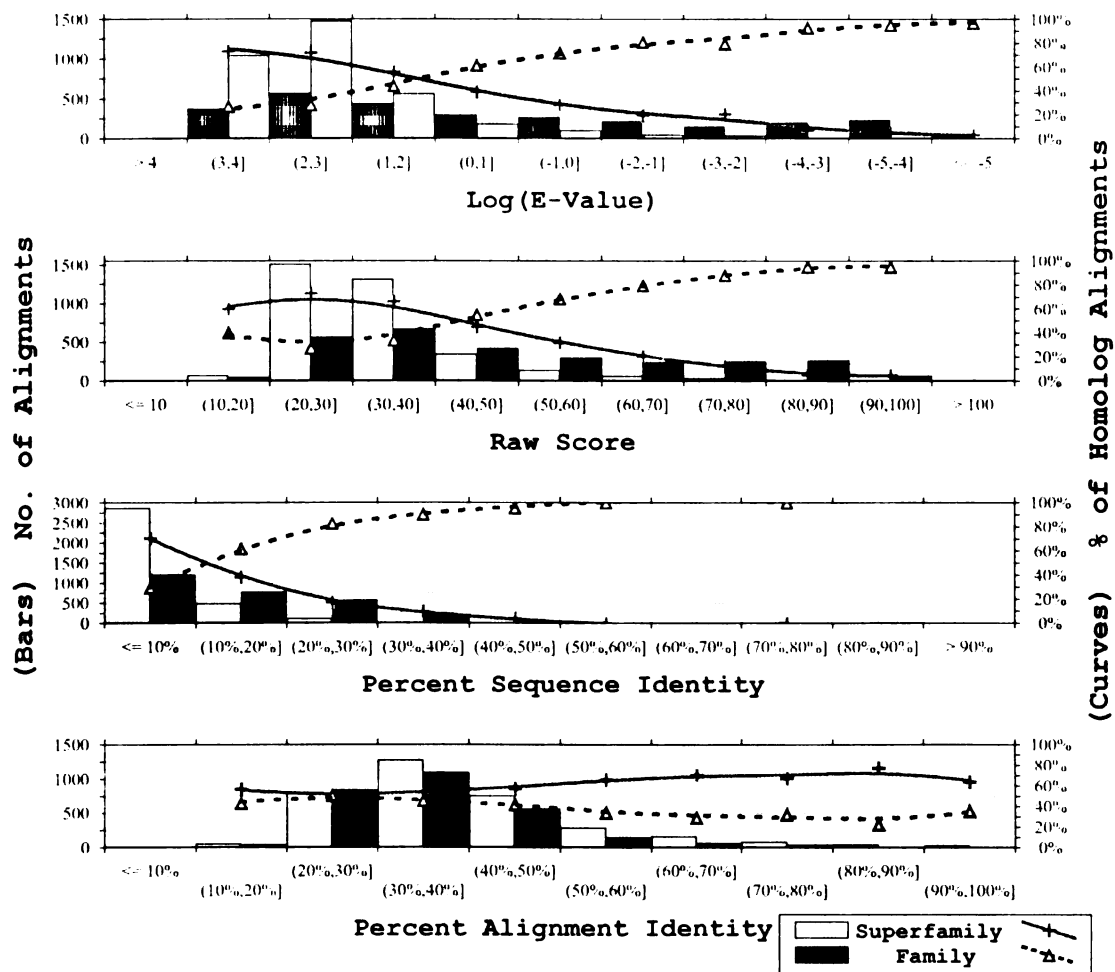
**Raw Score**

**Figure RSF1.** Raw score distributions for individual superfamilies and families.

Histograms for raw alignment scores for homolog alignments from the PDB90tU dataset. **A.** Alignments for homologs from individual superfamilies. Note that superfamily histograms exclude alignments between homologs in the same family. The largest family from each superfamily is shown in the graph directly below it. **B.** Individual family alignments.

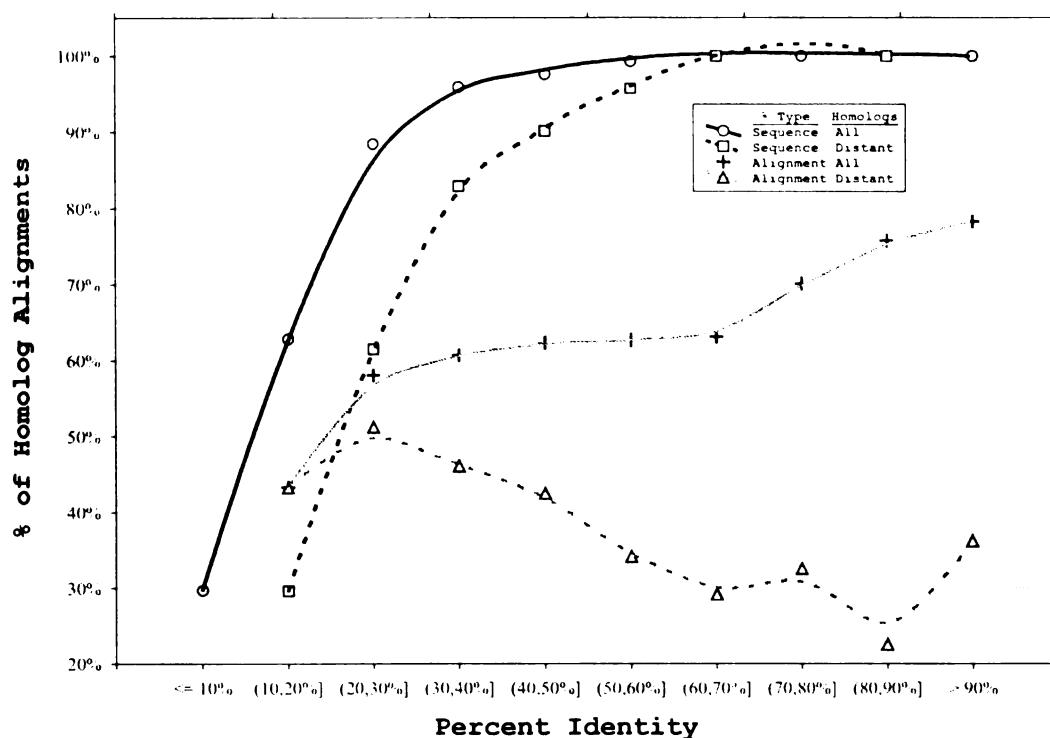


**Figure AID1.** Alignment length vs. identities for distant family and superfamily homologs. Scatterplots show all distant alignments (BLAST e-values  $< 10^{-5}$ ) from the PDB90tU dataset, categorized as family or superfamily homologs. Linear fits and variance explained by the fits ( $R^2$ ) are shown in the graph insets. Note that high-density areas in the plot may contain more data points than visually apparent, as points that overlap will appear as a single point. **A.** All distant superfamily (not-same-family) homologs. **B.** All distant family homologs. As discussed in (23) and Chapter III, non-homolog alignments score vs. alignment length relationships were similar to those of distant homolog alignments; and positives followed a linear relationships similar to those of identities.



**Figure ISF1.** Distribution of distant superfamily and family alignments for common alignment scores. All PDB90tU distant (BLAST e-values  $< 10^{-5}$ ) homolog alignments were categorized as family or superfamily alignments. Alignments for each category were binned by alignment scores as shown in the graph. Histogram bars show the number of alignments in each bin (left y-axis). Dark bars are for family alignments; light bars are for superfamily alignments. The points and curve fits show the percentage of family (triangles, dashed curve) or superfamily (crosses, solid curve) alignments in each bin. Percentages (right y-axis) were calculated as the number of (superfamily or family)

alignments in the corresponding x-axis score bin divided by the total number of homolog alignments (family plus superfamily) in that bin. When interpreting superfamily alignments as random alignments, the dashed curve illustrates the drop in the percentage of distant homolog alignments that may contain biological information at lower scores. The lack of a low-score family/superfamily trend in the percent alignment identity curves is consistent with that score being unrelated to alignment (or biological) similarity for distant homologs.



**Figure PFH1.** *Percent identity vs. fraction of same-family homologs.* This scatterplot expands on Figure ISF1 by comparing trends when all PDB90tU homolog alignments are included (“All”, solid lines), versus only distant PDB90tU homologs (“Distant”, dashed lines). The y-axis shows the fraction of homologs that are in the same family in a given dataset. The x-axis indicates both percent sequence identity and percent alignment identity (“% Type” = “Sequence” or “Alignment”, respectively). Percent sequence identity (dark lines) shows a steeper drop-off at lower scores when all, rather than only distant, homologs are considered. This drop-off occurs later than for only distant homologs, and appears at approximately 30-40% sequence identity. This drop-off in family alignments is similar to drop-offs in agreement between biological similarity and percent sequence identity as previously reported (see the recent studies (30, 40, 41) and

references therein). The trends shown for percent alignment identity (light lines) suggest that this score behaves very differently for distant and non-distant homologs. Raw score and e-value results for all homologs were largely unchanged from those for distant homologs (as in Figure ISF1), and are not shown.

## REFERENCES

1. **Abagyan, R. A., and S. Batalov.** 1997. Do aligned sequences share the same fold? *J Mol Biol.* **273**(1):355-68.
2. **Altschul, S. F.** 1998. Generalized affine gap costs for protein sequence alignment. *Proteins.* **32**(1):88-96.
3. **Altschul, S. F., M. S. Boguski, W. Gish, and J. C. Wootton.** 1994. Issues in searching molecular sequence databases. *Nat Genet.* **6**(2):119-29.
4. **Altschul, S. F., and W. Gish.** 1996. Local alignment statistics. *Methods Enzymol.* **266**:460-80.
5. **Benner, S. A., M. A. Cohen, and G. H. Gonnet.** 1993. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J Mol Biol.* **229**(4):1065-82.
6. **Brenner, S. E., C. Chothia, and T. J. Hubbard.** 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A.* **95**(11):6073-8.
7. **Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt.** 1978. A Model of Evolutionary Change in Proteins. *Atlas of Protein Sequence and Structure*:345-352.
8. **Domingues, F. S., P. Lackner, A. Andreeva, and M. J. Sippl.** 2000. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J Mol Biol.* **297**(4):1003-13.
9. **Doolittle, R. F.** 1986. *Of urfs and orfs : a primer on how to analyze derived amino acid sequences.* University Science Books, Mill Valley, CA.



10. **Durbin, R., S. Eddy, A. Krogh, and G. Mitchison.** 1998. Biological sequence analysis : probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, UK New York.
11. **Evans, M., N. A. J. Hastings, and J. B. Peacock.** 1993. Statistical distributions. 2nd / ed. J. Wiley, New York.
12. **Gotoh, O.** 1999. Multiple sequence alignment: algorithms and applications. *Adv Biophys.* **36**:159-206.
13. **Gumbel, E. J.** 1958. Statistics of extremes. Columbia University Press, New York,.
14. **Henikoff, S., and J. G. Henikoff.** 2000. Amino acid substitution matrices. *Adv Protein Chem.* **54**:73-97.
15. **Henikoff, S., and J. G. Henikoff.** 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A.* **89**(22):10915-9.
16. **Holmes, I., and R. Durbin.** 1998. Dynamic programming alignment accuracy. *J Comput Biol.* **5**(3):493-504.
17. **Jones, D. T., W. R. Taylor, and J. M. Thornton.** 1994. A mutation data matrix for transmembrane proteins. *FEBS Lett.* **339**(3):269-75.
18. **Lindahl, E., and A. Elofsson.** 2000. Identification of related proteins on family, superfamily and fold level. *J Mol Biol.* **295**(3):613-25.
19. **Lo Conte, L., B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin, and C. Chothia.** 2000. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* **28**(1):257-9.

20. **McNeil, K. A., I. Newman, and F. J. Kelly.** 1996. Testing research hypotheses with the general linear model. Southern Illinois University Press, Carbondale.
21. **Muller, A., R. M. MacCallum, and M. J. Sternberg.** 1999. Benchmarking PSI-BLAST in genome annotation. *J Mol Biol.* **293**(5):1257-71.
22. **Naor, D., and D. L. Brutlag.** 1994. On near-optimal alignments of biological sequences. *J Comput Biol.* **1**(4):349-66.
23. **Otillar, R. P., M. R. Segal, and C. A. Hunt.** 2001. Empirical Analysis of Percentage Similarity Measures Commonly Used for Understanding Distant Alignments. Submitted.
24. **Park, J., K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia.** 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol.* **284**(4):1201-10.
25. **Pearson, W. R.** 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol.* **132**:185-219.
26. **Reeck, G. R., C. de Haen, D. C. Teller, R. F. Doolittle, W. M. Fitch, R. E. Dickerson, P. Chambon, A. D. McLachlan, E. Margoliash, T. H. Jukes, et al.** 1987. "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it [letter]. *Cell.* **50**(5):667.
27. **Rost, B.** 1999. Twilight zone of protein sequence alignments. *Protein Eng.* **12**(2):85-94.
28. **Sander, C., and R. Schneider.** 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins.* **9**(1):56-68.

29. **Saqi, M. A., R. B. Russell, and M. J. Sternberg.** 1998. Misleading local sequence alignments: implications for comparative protein modelling. *Protein Eng.* **11(8):**627-30.
30. **Sauder, J. M., J. W. Arthur, and R. L. Dunbrack, Jr.** 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins.* **40(1):**6-22.
31. **Schaffer, A. A., Y. I. Wolf, C. P. Ponting, E. V. Koonin, L. Aravind, and S. F. Altschul.** 1999. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics.* **15(12):**1000-11.
32. **Sellers, P.** 1974. On the Theory and Computation of Evolutionary Distances. *J. Appl. Math (Siam).* **26(4):**787-793.
33. **Smith, T. F., and M. S. Waterman.** 1981. Identification of common molecular subsequences. *J Mol Biol.* **147(1):**195-7.
34. **Thompson, J. D., F. Plewniak, and O. Poch.** 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* **27(13):**2682-90.
35. **Venables, W. N., and B. D. Ripley.** 1999. *Modern applied statistics with S-PLUS*, 3rd ed. Springer, New York.
36. **Vingron, M.** 1996. Near-optimal sequence alignment. *Curr Opin Struct Biol.* **6(3):**346-52.

37. **Vogt, G., T. Etzold, and P. Argos.** 1995. An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *J Mol Biol.* **249**(4):816-31.
38. **Waterman, M.** 1983. Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. *Proc. Natl. Acad. Sci. USA.* **80**:3123-3124.
39. **Waterman, M. S.** 1995. *Introduction to computational biology : maps, sequences and genomes*, 1st ed. Chapman & Hall, London ; New York.
40. **Wilson, C. A., J. Kreychman, and M. Gerstein.** 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.
41. **Wood, T. C., and W. R. Pearson.** 1999. Evolution of protein sequences and structures. *J Mol Biol.* **291**(4):977-95.

## **ABSTRACT**

### **Motivation:**

Functional genomics, understanding the function of genes and the organization of genomes, is central to life science research. Comparative analysis by sequence alignment is the most rapid and widely used tool in functional genomics. Using current methods, sequence comparison methods are unable to reliably detect common ancestry, homology, in 60%-80% of cases where structural methods have already shown that common ancestry exists. Towards improving sequence-analysis methods for functional genomics, this work explored the empirical nature of the raw scores and statistical values that arise between homologous proteins of little or no statistically significant sequence similarity.

### **Experiments:**

We analyzed six hundred and forty thousand pairwise sequence alignments from an all-against-all alignment of the SCOP database; about ten thousand of these were alignments between homologs.

### **Results:**

To better understand the evolutionary distance when homology can no longer be reliably detected, proteins were broken out into family and superfamily comparisons. Consistent with previous reports, we found that 61% of family comparisons and only 4% of superfamily relationships could be detected above a 1% confidence cutoff. Surprisingly, we observed that the raw alignment scores of superfamily relationships were well described by an extreme value distribution, and that family relationships followed a mixture distribution consistent with an extreme-value component for low-scoring alignments. It is well known that alignments are increasingly random for low-scoring

homologs. However, our results extended this intuitive understanding, quantitatively suggesting that the raw scores of distantly related proteins were essentially random numbers, dominated by non-biological alignments much like those observed between unrelated proteins.

## INTRODUCTION

The recently completed drafts of the Human Genome sequence (13, 26) have reaffirmed our need to understand the functional and evolutionary relationships between genes, and between genomes. Alignment and comparative analysis, of each gene's protein sequence with previously studied sequences, is the principal method used for functional analysis of genes by recent genome projects (1, 7, 13, 26).

Common ancestry, homology, is perhaps the most basic form of biological similarity between genes captured by sequence comparison (e.g. (6, 9, 14)). Current methods for detecting homology are unable to detect biological homology in 60-75% of the cases from representative sets of proteins where common homology is known to exist by structural (i.e. non-sequence) comparison methods (6, 14, 16).

Because many biological relationships cannot be detected by sequence comparison methods, only 40-60% of newly sequenced genes can be annotated as homologs to gene families or superfamilies that allow even the most rudimentary functional classification (1, 7, 13, 26).

Several recent reports have analyzed the biological information (8, 22, 28) and homology detection performance (6, 14, 21) of pairwise sequence analysis algorithms for distantly related sequences. The most recent of these reports have highlighted the need to discern when alignments pair biologically conserved residues, noted as alignment 'quality' or biological 'correctness', (8, 22) and to distinguish the ability to detect homology at the family versus the (more evolutionarily distant) superfamily level (14, 22).

Superfamily relationships are much more difficult to detect and align correctly than family-level relationships. Homology can be reliably detected between sequences in the same family about 40% of the time, whereas homology can only be detected about 4% of the time for sequences sharing only superfamily similarity (14). Sequences within the same superfamily but different families are known to generate much lower quality alignments than sequences in the same family, even when both alignments have the same similarity scores (22). These results are not entirely surprising, given that sequences in the same family must share substantial common biology or sequence similarity, whereas any two sequences sharing homology—at any evolutionary distance—are classified into the same superfamily (15). As most sequences' homology relationships appear to be distant rather than close (see Chapter III and (6, 14, 17, 22)), understanding the nature of sequence statistics and alignments at the family and superfamily level appears critical to improving homology detection and the functional annotation of genes and genomes.

To better understand the biological signal in comparisons of distantly related proteins, we analyzed the alignment scores between sequences whose family and superfamily evolutionary relationships (or lack thereof) are known *a priori* by non-sequence methods. The alignments analyzed here, the PDB90tU dataset (see Chapter III and (17)), consisted of an all-against all alignment of the SCOP database (6, 15). Alignments were calculated using the NCBI Gapped BLAST algorithm with default scoring parameters and a permissive reporting cutoff, so that alignments even extremely distantly related sequences would not be filtered from BLAST's output. In all, 642,924 alignments including 9,314 homolog alignments (5,730 family- and 3,584 superfamily- alignments) were analyzed.



We extended earlier studies by examining the raw-alignment-score distributions of very-low-significance homologs. In breaking these down into family- and superfamily-distance alignments, we quantified evidence that the alignments of distantly related proteins were random and contained little or no biological signal. By characterizing the biological signal that remained to be captured by future methods, we gained new predictive insights into the barriers to detecting remote homologies.

## **METHODS**

### *Sequences, Sequence Databases, and Homology*

We used PDB90tU dataset of 642,924 alignments from an all-against-all alignment of the 1,937 PDB90t sequence database, both described in Chapter III and (17). Briefly, these sequences are a slightly reduced version of the SCOP PDB90D database of (6), with extremely short sequences and one highly over represented family removed. The alignments were calculated with the widely used gapped BLAST algorithm, using the default alignment scoring parameters (BLOSUM62 scoring matrix, -11/-1 gap open/extend penalty) and an extremely permissive alignment reporting cutoff (“Expectation value” cutoff = 10.000) to allow reporting of even very low scoring alignments. Alignments of a sequence to itself were removed from PDB90tU. PDB90tU contains at most one gapped alignment for any pair of sequences; when BLAST results reported an alignment between two sequences in both orders, i.e. “A↔B” and “B↔A”, a single alignment was kept with priority given to the higher scoring alignment. ‘Homology’ was taken as two sequences sharing the same SCOP class, fold, and superfamily. ‘Superfamily’ similarity was taken as two sequences sharing the same SCOP class, fold, and superfamily, but different SCOP families.

### *Measures and Formulas on Alignments*

All alignment parameters, including ‘e-values’ and Smith-waterman ‘raw scores’, were parsed directly from BLAST output and included in the PDB90tU database as described in Chapter III and (17).

To allow direct comparison of extreme value parameters and raw scores, the ‘location’ and ‘scale’ parameters commonly use in statistics were used in our extreme value fits

rather than the Karlin-Altschul representation of ‘K’ and ‘lambda’ (2, 12). To allow direct interpretation of the role of raw scores in the extreme-value process, fits were performed without corrections for the lengths of the aligned sequences. For extreme value curve fits, we note that

$$\text{extreme}(x, u, l) = \frac{1}{l} \exp\left(\frac{-(x-u)}{l}\right) \exp\left(\exp\left(\frac{-(x-u)}{l}\right)\right) \quad \text{for raw score } x, \text{ location } u, \text{ and}$$

scale  $l$  (c.f. (2, 11, 23)).

## RESULTS

### *Tabulation of Family, Superfamily Homology Detection*

A breakdown of homolog detection using statistical significance e-values for intra-family (‘family’) alignments and inter-family-homolog (‘superfamily’) alignments is summarized in Table BD1.

Consistent with previous results ((14)), we observed that the transition from family to superfamily relationships is coincident with an almost complete breakdown in the ability to detect homology with pairwise alignments. We noted that many family (15%) and superfamily (40%) alignments were not reported by BLAST at all, even at our extremely permissive cutoff for reporting alignments (BLAST e-value  $\geq 10,000$ ; see Methods). Overall, 48% of family- and 98% of superfamily- homolog pairs were not aligned with enough sequence similarity to reach statistical significance; i.e. the biological relationship between these homologs would not have been detected using their alignment scores. Most (62%) of the *reported* family-level alignments had statistically significant sequence similarity, whereas few (4%) reported superfamily-level alignments did, similar to the findings of (14). Consistent with previous reports (e.g. (6)), the combined family and

superfamily alignments achieved statistical significance for only 29% of the possible pairwise biological relationships.

#### *Comparison of Superfamily, Family Alignment Score Distributions*

To further explore the difficulties in detecting biological similarity across the family-superfamily barrier, we examined the raw scores that determined which of the many possible alignments between homologs was ranked ‘optimal’ and hence reported by our alignment software, BLAST. As (Figure IF1) shows, alignments of sequences within the same family follow a broad distribution of raw scores with several smaller peaks at the low end of the raw-score range. In stark contrast, superfamily alignments were almost exclusively low-scoring and appeared to be concentrated around a single peak at the low end of the score range.

#### *Comparison of Superfamily and Non-Homolog Score Distributions*

The single dominant peak in raw scores of distantly related alignments suggested that inter-family alignments might be driven by a single random process, rather than by capturing properties specific to each superfamily. Alignment scores for randomly generated sequences (and non-homologous sequences) stem from a single random alignment process: maximizing alignment score without constraints due to common biological ancestry (see (2) and references therein). The biological accuracy of homolog alignments is known to decrease at lower alignment scores (10, 22, 27). To investigate similarities between distant homolog alignments and non-homolog alignments, we fit extreme value distributions to the superfamily alignments and non-homolog alignments in our data (Figure ES1). The superfamily homolog alignments’ extreme value fit had a slightly higher location (raw score of 28.9 versus 26.0) and a somewhat broader distance

scale (5.8 versus 4.4) than the non-homologs (Figure ES1). This location difference of 2.9 raw score points is less than the raw score increase from one additional residue match per alignment. (The lowest residue-pair match is score +4, e.g. Serine-Serine, for the BLOSUM62 scoring matrix used in PDB90tU alignments). The homolog alignments also appeared slightly heavy in the right-hand tail (e.g. raw score 50-80, Figure ES1), whereas the non-homologs appeared somewhat heavier in the left-hand tail (raw score 10-20, Figure ES1). Despite these differences, the variance explained by these fits indicated that extreme value distributions explained our superfamily and non-homolog alignments similarly well ( $R^2=99.1\%$  and  $99.4\%$ , respectively).

#### *Single Family Score Distributions*

To investigate the nature of alignment accuracy at family-level evolutionary distances, we examined raw-score core distributions for the four families with the largest number of PDB90tU alignments (Figure RF1).

For the Eukaryotic Protease family, one large peak appeared visually to have a location and overall shape similar to that of unrelated sequences' alignments (compare Figures RF1 and ES1). This large peak appeared to be composed of smaller score-cluster peaks located around raw scores of 28, 34, and 39. The score distributions for the Animal Virus Protein ('AVP'), Globin, and C1 Set Domain ('C1') families were more evenly dispersed; with Globins and C1s having density concentrated in the lower score range (0-125). It was unclear if the visual peaks in the raw score distributions for the AVP, Globin, and C1 families were justifiably explained as clusters of alignment scores. Overall, each family's alignment scores appeared to follow different distributions; proposed implications of these distributions are addressed in the Discussion.

## DISCUSSION

### *Superfamily Homologs Appeared Randomly Aligned*

It is widely held that as evolutionary distances increase, the quality of alignments goes down (see (22, 27)). Empirical and theoretical analyses support this conclusion. Empirically, residue alignments based on three-dimensional structure comparison are used as a gold standard for evaluation of sequence-based alignment accuracy. Family and superfamily alignments based on sequence comparison often fail to agree with the structural alignment for when the alignment score is low (e.g. (22)). Theoretically, if a sub-region of a potential alignment is biologically correct, but the overall alignment has a lower score than a different—and non-biological—alignment, then the non-biological alignment will be reported as the ‘optimal’ (see (27)).

For PDB90tU data, the median score for superfamily sequences’ alignments (raw score = 31) was close to the median score for unrelated sequences’ alignments (raw score = 27). The difference of 4 raw score points is similar to the score from adding a single additional residue match (e.g. for the BLOSUM62 scoring matrix used in calculating PDB90tU alignments, Alanine-Alanine = +4; Threonine-Threonine = +5). These medians suggested that scores for non-biological alignments were often higher than biological or partly-biological alignments between PDB90tU superfamily homologs, and that many PDB90tU superfamily alignments may have been random or mostly random.

A key signature of alignments between unrelated sequences (and between randomly generated sequence) is that the alignments’ raw scores will follow an extreme value distribution (12). This signature was present in PDB90tU non-homolog alignments (Figure ES1). Surprisingly, the raw scores of superfamily alignments in our data followed

an Extreme-value distribution similar to that of non-homolog alignments (Figure ES1). An extreme value distribution explained both non-homolog- and superfamily-alignments' raw scores similarly well ( $R^2 = 99.4\%$  and  $99.1\%$ , respectively).

We concluded that the bulk of PDB90tU superfamily alignments were dominated by an alignment process similar to the one that occurs between non-homolog sequences. To wit, PDB90tU superfamily alignments generally appeared to have little or no bias towards pairing up residues that descended from the same ancestor residue or DNA codon.

#### *Low-scoring Family Homologs May Be Randomly Aligned*

Family alignments formed a peak that appeared visually consistent with an extreme-value distribution process for low scoring alignments (Figure IF1, raw score 20-50 range). We hypothesized that the family alignments were a mixture of biological and non-biological processes, resulting in a mixture distribution with an extreme-value peak dominating the lower raw scores and other (non-extreme-value) biological alignment processes dominating the higher raw scores. We were unable to quantify this hypothesis using Figure IF1, as the low score peak was broader on the right hand side than predicted by a single extreme-value process, and the peak not well enough separated from other peaks of unknown distributions to allow accurate single-distribution fits with available statistical tools. (We felt that hand-adjustment of the extreme value scale and location parameters to produce a visually appealing fit would add no new insights. As mixture distribution fitting is not widely used in the sequence-analysis literature, we note that current statistical tools and methods (4, 23, 25) are not well suited to fitting mixtures containing extreme value distributions that closely overlap with other, possibly unknown, distributions).

Several peaks appeared to be present in the lower-scoring range of family alignments (e.g. Figure IF1, raw score ranges 20-50, 75-95). We hypothesized that these peaks corresponded to random alignments for homologs within the same family.

Compared to superfamily homologs, family homologs have undergone fewer evolutionary events (e.g. insertion/deletions in loop regions, repeated mutations of each residue) that would cause random alignments to have the same scores as non-homolog sequences' alignments. Hence, the variation in statistics for low-scoring family alignments may be due to random alignments with higher score bias from biological similarities that do not depend on an accurate alignment, including overall residue bias, residue bias in overlapping secondary-structures, or common sequence lengths. This was expected to cause a relatively uniform increase in the score of random alignments between homologs.

Alternatively, a family's alignments with scores centered on a low-score peak might be anchored by a high-scoring core alignment segment that is biologically correct. Clusters of family members sharing the same conserved, correctly aligned subsegment would have raw alignment scores distributed around the same peak. This was expected to cause multiple peaks in raw score for each family's alignments, with each peak corresponding to a random, extreme-value-distributed alignment score plus a non-random biologically-correct-core-alignment score that shifted the center of a baseline extreme-value distribution rightward, to a higher score.

A striking difference between families in our analysis was the raw score distribution of the AVP family compared to the other individual families analyzed (Figure RF1). The AVP family appeared to support the conserved-core alignment similarity hypothesis, with



multiple peaks and an overall distribution similar to that of superfamily homologs and unrelated sequences. The lack of an apparent extreme-value peak for the other families was consistent with the conclusion that their alignment scores were driven by a different process than the random alignment observed between non-homologous sequences.

#### *Sequence Assays Use Non-Biological Alignments for Most Homologs*

An extreme value distribution explained 97% of the raw score for non-significant PDB90tU homolog alignments. The extreme value fit suggested that little or no biological signal was captured in approximately 5,268 of these alignments (Figure AH1, Table BD1). 3,383 homolog pairs had no PDB90tU alignment (Table BD1), i.e. they had no alignment calculated by BLAST that scored above an expectation cutoff of 10,000. In total, these 8,651 homology relationships represented over two thirds of all 12,697 homology relationships in the PDB90t sequences. Even if all family alignments were assumed biologically accurate, Figure ES1 suggests that approximately 3,316 superfamily alignments, and hence 53% of all PDB90tU homology relationships, did not receive alignment scores driven by biological residue pairings. Hence, for most PDB90tU homologs, the raw scores used in final e-values and statistical assays of sequence homology appeared to be based on non-biological alignments.

We noted that our observed 1% cutoff for an extreme value fit to all low-significance homolog alignments was close to the 1% cutoff for significant homology detection (Figure AH1). We propose that this may be causal, and that the primary difficulty in improving sequence comparison methods is that the distant alignments being analyzed are random and contain little or no signal of biological conservation. To wit, the problem

in developing assays for detecting remote homology may not overcome the ‘background noise’ of unrelated sequences’ high-scoring alignments. Rather, it appeared that most distant homolog alignments themselves are generally random, *are part of the background noise*, and have no signal of biological conservation to be detected.

Improving the biological signal with more ancestrally correct alignments between low-scoring homologs appeared a requirement before methods could be devised for scoring homologous relationships above the noise of non-homolog random alignments. An emphasis on understanding alignment quality, rather than statistics of optimal alignments, appeared indicated for reliably detecting distant evolutionary relationships (see (22, 24)).

#### *Our Results May Apply to Alignment Algorithms in General*

The data and results here for Smith-Waterman based alignments agreed with the accepted intuition that weak sequence conservation results in lower quality alignments *and* with data from previous reports. The PDB90tU alignment data have been analyzed as consistent with those reported using a wide range of full and approximate Smith-Waterman based algorithms, scoring schemes, and data sets (see Chapter III and (17)). Hence, we concluded that the analyses and results reported here were likely to apply to Smith-Waterman-based algorithms in general.

The quantifiably random nature of PDB90tU alignments may extend to other alignment methods, including structural alignment methods, hidden Markov models, and multiple-sequence profile methods (as proposed for relationships studied in Chapter III and (17)). The randomness in the distant alignments reported by an algorithm may be tested by comparing alignment scores for distantly related sequences versus unrelated sequences for a common distribution. If distant homolog alignments show a score distribution that is

the signature for alignments between unrelated sequences, as was seen for BLAST in Figure ES1, the biological information in those alignments should be suspect.

#### *Biological Applications and Better Alignments*

Without accurate scores, it is difficult to tell the difference between non-homolog and distant family relationships by alignment methods (e.g. (6, 9)). To our knowledge, we have reported the first observation that distantly related *homolog* alignments follow an extreme value distribution, and are thus likely to have little or no biological signal to be gleaned from their alignments. At the time of this writing, datasets with many distant homolog alignments comparable to those in PDB90tU remain commonly used in the development of new methods for remote homology detection or comparisons of distant biological conservation to other types of similarity (see Chapter III, (17) and references therein.) Going forward, we propose that an initial examination of sequence alignment score distributions could determine which subsets of a dataset's alignments are unlikely to contain sufficient biological signal to support new biological conclusions.

Towards extracting useful information from particular distant alignments, we suggest that extreme-value fits to homolog alignments (e.g. Figures ES1, AH1) may be used as weighting factors or probabilities indicating the likelihood that that an alignment between putative homologs would be randomly aligned. Similar pre-calculated fits are already widely used in homology assays and the calculation of BLAST e-values (2, 3). A weighting of alignments between known homologs could assign quantified confidence of an alignment's biological relevance in new meta-analysis procedures that combine multiple sources of information to draw functional genomics conclusions (26), like

intermediate sequence searches (18) or combining gene-expression, annotation text similarity, genomic context, and other measures of gene function (e.g. (5, 19, 20)).

Towards generating better alignment algorithms, we suggest that the results reported here may provide an assay for improving alignments of distant homologs without actually knowing the correct alignment. A decrease in the number of superfamily, or distantly related, homolog sequences following an extreme-value distribution should provide an assay for scores and algorithms that provide improved biological alignments. An advantage of aligning sequences and then analyzing their score distributions, rather than the exact residues paired, is that the correct alignment between distantly related sequences was not required to ascertain the distribution of distantly related sequences (e.g. Figures IF1, ES1, AH1). This approach is valid for algorithms and scores that keep an extreme-value maximum score for non-homolog sequences, but does not require three-dimensional structural comparisons or *a priori* knowledge of correct sequence alignments.

It has been noted that family alignments are more biologically informative than superfamily alignments, even at the same level of similarity (22). For families like the Eukaryotic Proteases, Globins, and C1s, this may be explained by the fact that their scores appeared dominated by a different (and presumably more biologically driven) alignment process than occurs in superfamily alignments and non-homolog alignments (see Figure RF1). Comparing families with scores in the same range, but with different distributions (e.g. Eukaryotic Proteases versus AVPs, Figure RF1), may yield new insights into the signatures in residue pairings that signify biological conservation rather than random-but-high-scoring homolog alignments. Future analyses contrasting family

and superfamily pairs chosen as comparable by non-sequence similarity methods—and *then* aligned, scored, and tested for residue-by-residue accuracy using sequence methods—may shed additional light on the relationship of evolutionary distance, biological conservation, and sequence-alignment similarity.

## **CONCLUSION**

A better understanding of alignments between distantly related sequences, and better tools for using sequences to measure distant biological relationships, is vital to advancing our understanding of the biomedical properties of genes, and genomes. It is widely understood that alignments of distantly homologs are less biologically correct than alignments of closely related sequences. Our data extended this understanding, providing quantitative evidence that distant homologs' alignments followed an extreme-value distribution, and that most of the residue pairings those alignments were largely random. Our data suggested that most homolog relationships were not evaluated using biologically meaningful relationships, and that the principal barrier to detecting remote homology was calculating better homolog alignments, rather than improving homolog alignment scores or lowering background noise from unrelated sequences. Because distant homolog alignments followed a raw score distribution much like that of unrelated sequences, we proposed that baseline fits of homolog alignment scores could provide assays for improving existing alignment methods and weighting statistics for combined-evidence tools like intermediate sequence searches. Powerful tools and scientific advances for using extreme-value fits of unrelated sequences have already provided functional characterization of 40-60% the genes in each newly sequenced genome.

Leveraging those works with extensions to distant-but-homologous alignments may enable characterization of the remainder.

**Table BD1.** *Breakdown of homolog**detection level by evolutionary distance.*

The non-percentage numbers in this 3x2

breakdown count all pairwise homology

relationships between the 1,937 PDB90t

sequences. Percentages indicate row,

column, or table-total percentages of the

homolog relationship counts. ‘Family’

evolutionary distance indicates two

Detection Level	Evolutionary Distance		Row Totals
	Superfamily	Family	
Not Reported	2,355	1,028	3,383
<i>Column %</i>	40%	15%	
<i>Row %</i>	70%	30%	
<i>Total %</i>	19%	8%	27%
Not Significant	3,435	2,205	5,640
<i>Column %</i>	58%	33%	
<i>Row %</i>	61%	39%	
<i>Total %</i>	27%	17%	44%
Significant	149	3,525	3,674
<i>Column %</i>	2%	52%	
<i>Row %</i>	4%	96%	
<i>Total %</i>	1%	28%	29%
Column Totals	5,939	6,758	12,697
<i>Total %</i>	47%	53%	

PDB90t sequences in the same SCOP family; ‘Superfamily’ distance indicates two

PDB90t sequences in the same SCOP superfamily but different SCOP families (as noted

in the Methods). A ‘Significant’ detection level indicates a sequence pair with a

PDB90tU alignment scoring below a 1% significance cutoff, i.e. with BLAST e-value &lt;

1%. ‘Not Significant’ indicates a PDB90tU alignment above that cutoff. ‘Not Reported’

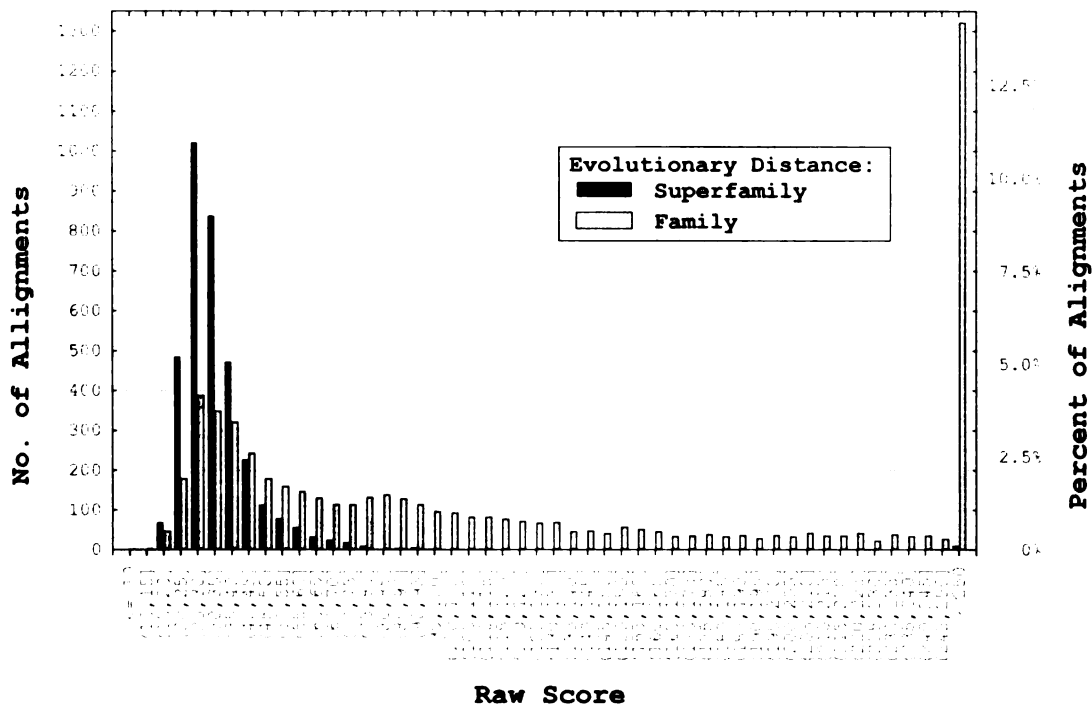
indicates no PDB90tU alignment exists between those two homologous sequences. Note

that, just as in the PDB90tU alignment set, Sequence A ⇔ Sequence B and Sequence B

⇔ Sequence A relationships are equivalent and do not count as two relationships, and

self relationships (i.e. Sequence A ⇔ Sequence A) are not included.

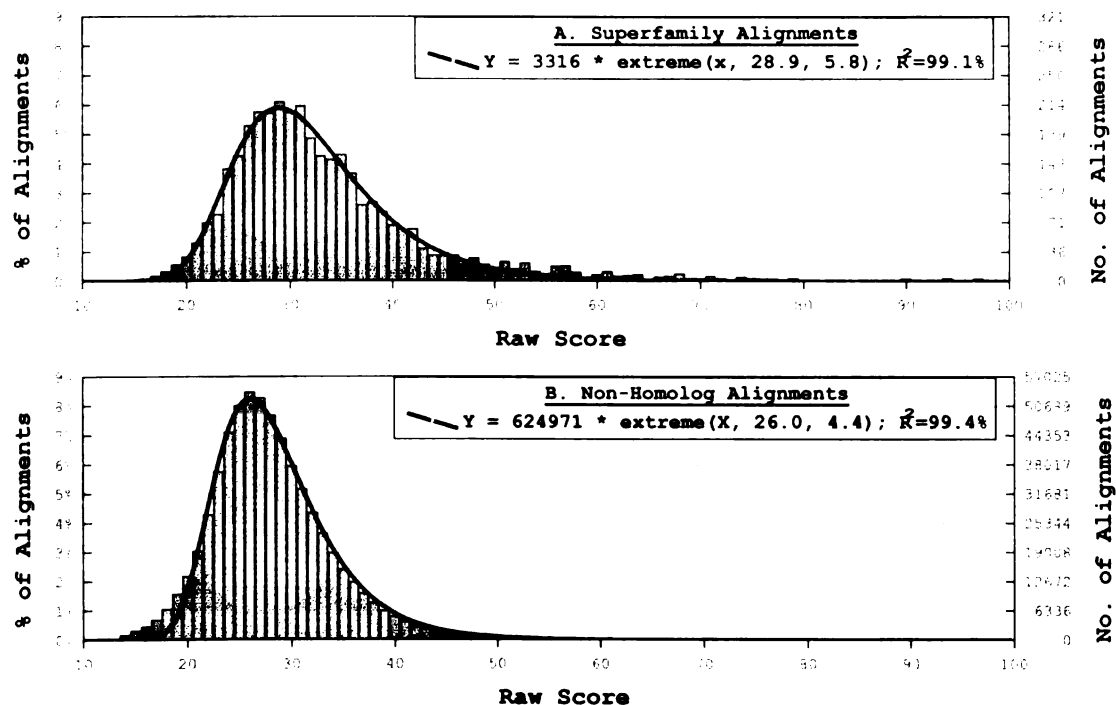
### Comparison of Family vs. Superfamily Alignment Score Distributions



**Figure IF1.** Histogram of the alignment raw-scores for all PDB90tU alignments between homologs, broken down by family and superfamily evolutionary distance. The ‘Percent of Alignments’ is calculated as a percent of the total number of PDB90tU homolog alignments (i.e. 9,314). Histogram bins width was five raw score points. Note that e-values are calculated from raw scores (see Methods) and the two are highly correlated. For PB90tU homologs, the minimum raw score for an alignment having a statistically significant e-value (i.e. e-value  $\leq 1\%$ ) was 64; the maximum raw score corresponding to a non-significant e-value (i.e. e-value  $> 1\%$ ) was 73.

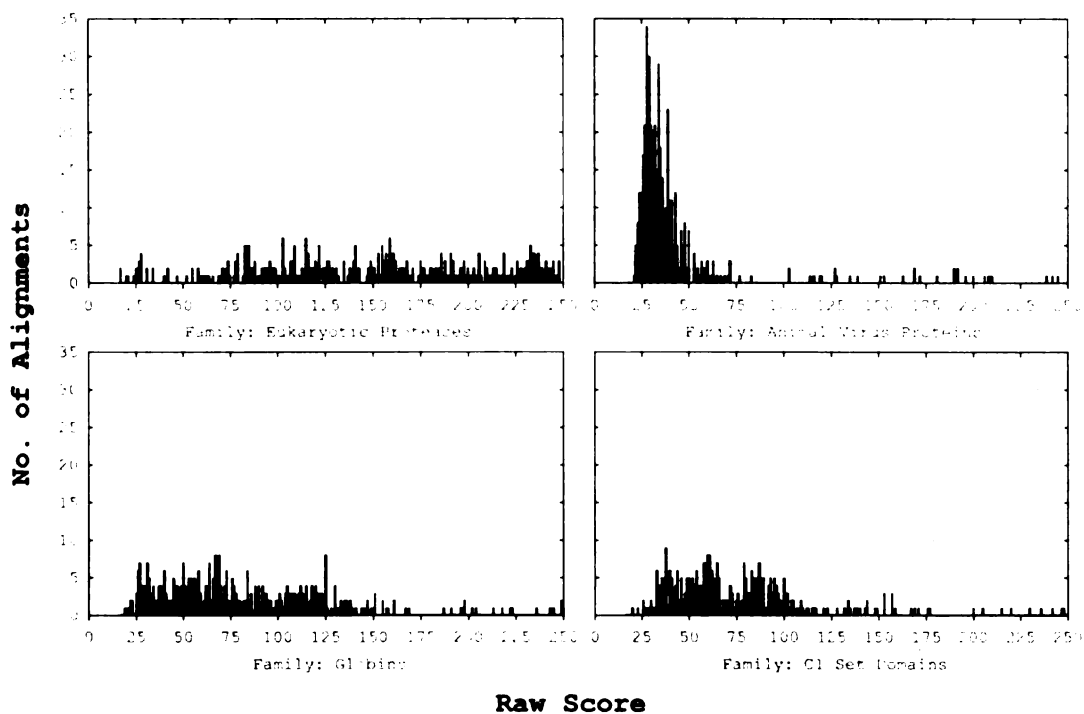


Extreme Value Fits to Superfamily and Non-Homolog Alignment Scores

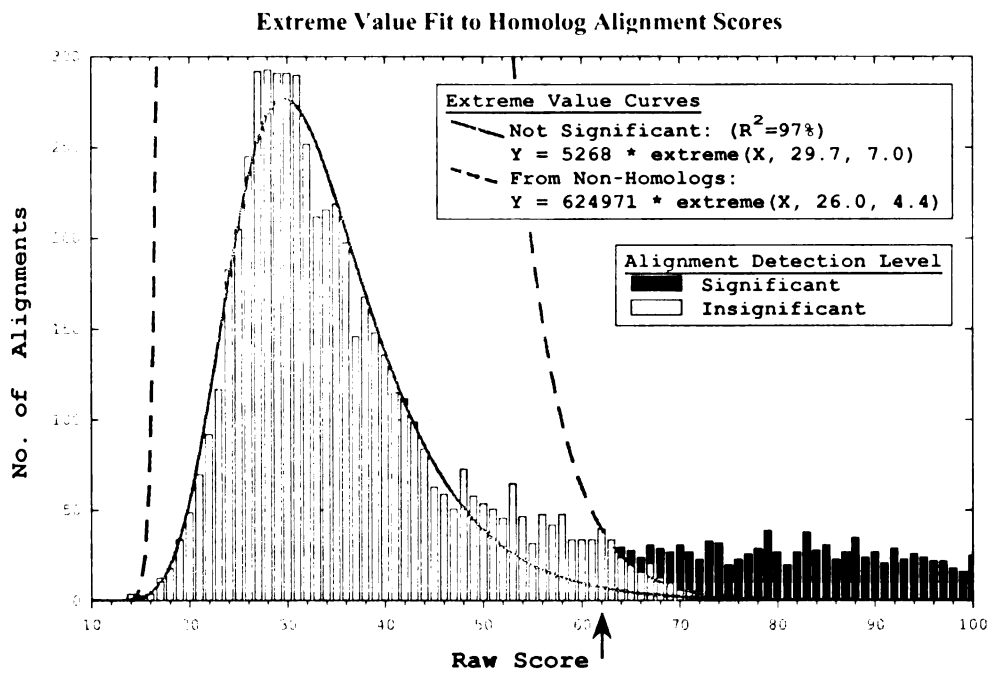


**Figure ES1.** Extreme value fits to superfamily and non-homolog alignment scores. Histograms and extreme value distribution curve fits for all PDB90tU **A.** superfamily alignments; **B.** non-homolog alignments. Fit data was taken as raw scores with histogram bin counts from this graph (bin width of one point raw score). Fit was performed by nonlinear estimation on  $\{Y = N * \text{extreme}(X, \text{location}, \text{scale})\}$  using quasi-Newton fitting.

### Examples of Alignment Score Distributions Within Individual Families



**Figure RF1.** *Examples of alignment score distributions within individual families.* Histograms for the four PDB90t families with the most PDB90tU alignments (bins width of one point raw score). Listed as (no. PDB90tU alignments, no. alignments below the shown 250 raw-score cutoff), they were: Eukaryotic Proteases (580, 371), Animal Virus Proteins (548, 454), Globins (496, 391), C1 Set Domains (361, 330). To provide a familiar reference point, we note that the C1 Set Domains are in the Immunoglobulin superfamily.



**Figure AH1.** Extreme value fit to homolog alignment scores. A stacked histogram of significant and non-significant scoring PDB90tU alignments (bin width was one point raw score). The solid curve extreme-value fit explained 97% of the variance in the raw score for the 5,640 non-significant (See Table BD1) PDB90tU alignments; this fit predicted that 5,268 alignments with similarly distributed raw scores would have been expected as results of a random extreme-value process. Note that an extreme value fit to all shown data yielded  $Y = 5305 * \text{extreme}(X, 29.7, 7.1)$ ,  $R^2=92\%$ . (Fit data was taken as raw scores with histogram bin counts from this graph; fit was performed by nonlinear estimation on  $\{Y = N * \text{extreme}(X, \text{location}, \text{scale})\}$  using quasi-Newton fitting.) The black arrow indicates the 1% confidence cutoff for the non-significant homologs' extreme value fit (solid curve at raw score = 62). The dashed curve indicates the extreme value fit to non-homolog alignments (taken from Figure ES1).

## REFERENCES

1. **Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, *et al.*** 2000. The genome sequence of *Drosophila melanogaster*. *Science*. **287**(5461):2185-95.
2. **Altschul, S. F., and W. Gish.** 1996. Local alignment statistics. *Methods Enzymol.* **266**:460-80.
3. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**(17):3389-402.
4. **Böhning, D.** 1999. Computer-assisted analysis of mixtures and applications : meta-analysis, disease mapping, and others. Chapman & Hall/CRC, Boca Raton, Fla.
5. **Bork, P., and E. V. Koonin.** 1998. Predicting functions from protein sequences--where are the bottlenecks? *Nat Genet.* **18**(4):313-8.
6. **Brenner, S. E., C. Chothia, and T. J. Hubbard.** 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A.* **95**(11):6073-8.
7. **Consortium, T. C. e. S.** 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science.* **282**(5396):2012-8.
8. **Domingues, F. S., P. Lackner, A. Andreeva, and M. J. Sippl.** 2000. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J Mol Biol.* **297**(4):1003-13.

9. **Doolittle, R. F.** 1986. Of urfs and orfs : a primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley, CA.
10. **Durbin, R., S. Eddy, A. Krogh, and G. Mitchison.** 1998. Biological sequence analysis : probabalistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, UK New York.
11. **Evans, M., N. A. J. Hastings, and J. B. Peacock.** 1993. Statistical distributions, 2nd / ed. J. Wiley, New York.
12. **Karlin, S., and S. F. Altschul.** 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A. **87**(6):2264-8.
13. **Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al.** 2001. Initial sequencing and analysis of the human genome. Nature. **409**(6822):860-921.
14. **Lindahl, E., and A. Elofsson.** 2000. Identification of related proteins on family, superfamily and fold level. J Mol Biol. **295**(3):613-25.
15. **Lo Conte, L., B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin, and C. Chothia.** 2000. SCOP: a structural classification of proteins database. Nucleic Acids Res. **28**(1):257-9.
16. **Muller, A., R. M. MacCallum, and M. J. Sternberg.** 1999. Benchmarking PSI-BLAST in genome annotation. J Mol Biol. **293**(5):1257-71.
17. **Otillar, R. P., M. R. Segal, and C. A. Hunt.** 2001. Empirical Analysis of Percentage Similarity Measures Commonly Used for Understanding Distant Alignments. Submitted.

18. **Park, J., K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia.** 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol.* **284(4):**1201-10.
19. **Pellegrini, M.** 2001. Computational methods for protein function analysis. *Curr Opin Chem Biol.* **5(1):**46-50.
20. **Rigoutsos, I., A. Floratos, L. Parida, Y. Gao, and D. Platt.** 2000. The emergence of pattern discovery techniques in computational biology. *Metab Eng.* **2(3):**159-77.
21. **Rost, B.** 1999. Twilight zone of protein sequence alignments. *Protein Eng.* **12(2):**85-94.
22. **Sauder, J. M., J. W. Arthur, and R. L. Dunbrack, Jr.** 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins.* **40(1):**6-22.
23. **Statsoft.** 1999. *Statistica*, 5.5 ed. Statsoft, Inc., 2300 East 14th Street, Tulsa, OK 74104.
24. **Taylor, W. R.** 1997. Multiple sequence threading: an analysis of alignment quality and stability. *J Mol Biol.* **269(5):**902-43.
25. **Venables, W. N., and B. D. Ripley.** 1999. *Modern applied statistics with S-PLUS*, 3rd ed. Springer, New York.
26. **Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al.** 2001. The sequence of the human genome. *Science.* **291(5507):**1304-51.

27. **Vingron, M.** 1996. Near-optimal sequence alignment. *Curr Opin Struct Biol.* **6(3):346-52.**
28. **Wilson, C. A., J. Kreychman, and M. Gerstein.** 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.**

**CHAPTER III: EMPIRICAL ANALYSIS OF PERCENTAGE SIMILARITY  
MEASURES COMMONLY USED FOR UNDERSTANDING DISTANT ALIGNMENTS**



## **ABSTRACT**

### **Motivation**

Sequence alignment scores such as percent similarity and percent identity are widely used as measures of sequence similarity and biological conservation between genes, even at low levels of similarity. These measures are less accurate for lower scoring alignments but are viewed as intuitive and straightforward to interpret. In quantitative studies, they routinely provide the sequence similarity “x-axis” for comparisons that include alignments of marginal- to non-significant (‘weak’) extreme value similarity statistics. This report explored the biological meaning of percentage-based scores on weak similarity alignments.

### **Results**

I calculated 642,924 alignments for Structural Classification Of Proteins database sequences, using the BLAST approximate Smith-Waterman program, and analyzed the redundancy of alignment length, number of identical residues, and number of positive-scoring residues. 99.9% of sequence pairs below the “twilight zone” sequence similarity cutoff had weak similarity. For weak similarity sequences, a linear function of alignment length strongly predicted the number of identical ( $R^2 > 81\%$ ) and positive-scoring ( $R^2 > 91\%$ ) residues in each alignment, in homolog and non-homolog alignments. Hence, for comparing weak similarity sequences, (1) percent alignment identity, the ratio of identical residues to alignment length, was not meaningful; (2) percent sequence identity, the ratio of identical residues to sequence length, was equivalent to the more intuitive “fraction aligned” measure; (3) ratios of similarity to alignment length or sequence

length, including similarity as number of positive-scoring residue pairs, were respectively non-meaningful or replaceable by more intuitive “fraction aligned” measures; (4) the HSSP boundary, the currently accepted relationship between alignment length and percent alignment identity, was reproducible as an artifact of the high correlation between alignment length and number of identical residues for non-homologs. These linear relationships and conclusions have not been previously reported; however, the agreement of my alignments with data from other publications (for programs including BLAST, SSEARCH, and MaxHom) suggested that the results reported here are general properties of Smith-Waterman based algorithms. Implications of these results for a broader range of alignment methods, including structural alignments, PSI-BLAST, and Hidden Markov Models are discussed.

## INTRODUCTION

A fundamental goal in life science research is to understand the biomedical properties and evolutionary relationships of genes and genomes. Because of the tremendous cost and time required to study each individual gene or hypothesized relationship using wet-lab measurements, genes' protein and DNA sequences are routinely compared for signatures indicating the likelihood of shared ancestry, structure, or biomedical function. When a measure of sequence-similarity can be formalized into a computational algorithm, it may be translated into software and used as a conduit of information between thousands to millions of previously studied gene sequences. Examples of successful computational gene comparisons include the rooting of the tree of life (19), the first established connection between oncogenes and normal growth factors (11), and the mapping and analysis of the human genome (28).

Comparing sequences basically requires two steps: finding high-scoring alignments between the sequences, and then ascertaining what biological properties—if any—those alignments indicate are likely to be shared. Full- and approximate- Smith-Waterman search algorithms ('SW's) efficiently calculate many (or all, respectively) possible ways of pairing residues between two sequences and report the highest-scoring alignments as optimal or near-optimal. SWs use alignment scores based on modern parameterizations (4, 9, 13, 15, 20, 33) of Sellers' 1974 mutation plus "indel" (insertion/deletion) measure of sequence evolution (29), and have been built into sophisticated, feature-rich software suites (e.g. the FASTA/SSEARCH suite (21), the BLAST suite (5)) by leading academic scientists. Decades of research on sequence alignments have provided accurate extreme-value statistics ('EVS') of SW scores for 'non-homologous' sequences (4, 15), i.e. those

with no common ancestor. EVS scores are now an accurate positive assay for common ancestry ('homology') between genes, with a false positive error rate below 1% (6). Over the last decade, functional and evolutionary assays on genes using the above-mentioned suites and statistics have become a standard protocol for analyzing all new genes and genomes, resulting in basic functional characterization for roughly 60% of newly sequenced genes (e.g. (1, 8)).

Once an alignment between two sequences is calculated, measures of percent identity are widely used to measure similarity between sequences. (percent alignment identity =  $I/L$ ; percent sequence identity =  $I/L_{max}$ ;  $I$  = identities, the number of identical-residue pairs in the alignment;  $L$  = alignment length, the number of non-gap residue pairs aligned; and  $L_{max}$  = length of the shortest sequence, the maximum possible alignment length). As a quantitative "x-axis" of sequence similarity, these percent identity measures remain used to judge homologs' biological similarity, particularly for statistical comparison to protein structure similarity (7, 24, 26, 34). Analyses using percent identity include alignments in marginal, non-significant, or unreported EVS score ranges (e.g. (6, 22, 24, 26)). Heuristic rules and percent identity cutoffs are effective positive assays for homology (10, 22, 24) and are reported (22) (and disputed (6)) as useful homology assays when EVS significance is marginal or non-existent.

To further the development of techniques for studying the 40% of genes, which remain uncharacterized by current sequence alignment methods, I analyzed relationships of alignment length, identity, and EVS scores for all pairwise alignments of Structural Classification of Proteins ('SCOP')(17) sequences generated by the NCBI gapped-BLAST (5) approximate SW software. SCOP provides prior knowledge of which

sequence pairs are homologous, based on non-sequence methods (17), and is established as a published standard for alignment evaluation with SW programs including BLAST (6, 26). BLAST results are essentially equivalent to those of other SW programs (3), and BLAST allowed computationally tractable calculation of numerous low-scoring alignments. Moreover, improving biological interpretations of BLAST alignments and scores will directly impact the work of its many scientific users in whole genome analysis, single-sequence searches by wet-lab biologists, and bioinformatics algorithm development (6, 22, 26).

Unlike the methods of previous empirical studies addressing identity and alignment length of biological sequences (6, 22, 25), this work focused on the direct relationship between alignment length and number of identities rather than using their ratios to measure sequence similarity; kept all alignments down to an extremely permissive cutoff (BLAST EVS 'expect' cutoff = 10,000); and separately analyzed 'strong' similarity (expect  $\leq$  0.00001) and 'weak' similarity (expect  $>$  0.00001) alignments. The strong/weak similarity partition was chosen to separate alignments that indicated high likelihood of the basic biological similarity of common ancestry (strong similarity) from those where all biological similarity is unclear (weak similarity). Weakly similar genes are exactly those that cannot be reliably compared for genomic annotation using SW tools; hence, the empirical strong/weak partition was used, rather than the traditional 'twilight zone' 25% sequence identity cutoff derived from evolutionary mutation distance (10). Additionally, I observed that nearly all alignments (>99.9%) below 25% sequence identity were weakly similar, indicating that analyses for weakly similar alignments included most alignments below the twilight zone cutoff.

In contrast to previous studies' focus on indirectly relating alignment length and percent alignment identity through a cutoff boundary curve distinguishing homolog and non-homolog alignments (2, 6, 22, 24), my analysis indicated that, *for weak similarity SW alignments*, the number of identities ( $I$ ) was well predicted as a linear function of the length of the alignment ( $L$ ) plus a small error term ( $\varepsilon$ ) (Figure 1). Hence, for weak similarity SW alignments, the percent alignment identity ( $I/L$ ) was effectively dividing alignment length by itself, i.e.  $L/L$ . The broad range of percent alignment identity widely observed in short alignments was due to the increasing influence of the error term ( $\varepsilon$ ) contribution ( $\varepsilon/L$ ) as the alignment length ( $L$ ) became small. Similarly, percent sequence identity ( $I/L_{max}$ ) was effectively equivalent to alignment length divided by sequence length ( $L/L_{max}$ ) plus a small error contribution ( $\varepsilon/L_{max}$ ).

These results indicated that common sequence-similarity measures of SW alignment percent identity are not appropriate for analysis of weakly similar sequences, including most 'distant' and twilight zone homologs. For these sequences, percent alignment identity largely reflected an inflated error term, whereas the information in percent sequence identity is more accurately reflected by fraction aligned as used in (26). Thus, by way of example, plots using percent alignment identity of SW alignments (e.g. (22, 24)) are unlikely to be comparing actual differences in biological- or sequence- similarity for their weakly-similar-sequence data points, and are potentially misleading for studying distantly related proteins.

## METHODS

### *The Sequence Database*

For comparison to a thorough and thoughtful published work on pairwise sequence comparison, I used a slightly reduced version of the SCOP PDB90D sequence database of (6) downloaded from <http://sss.berkeley.edu/db/scopseq/sdqib90-1.35.seq.fa> (2,079 sequences). As BLAST's edge-correction terms are valid for sequences longer than “1/K” (4, 16), with “gapped K”=0.047 in this study, sequences of length < 25 residues were removed (18 removed). The Immunoglobulin V set domain family was over represented and likely to introduce bias (26) and was removed; the Immunoglobulin V set domain family had 124 members in the original PDB90-B dataset, compared to 39, 36, and 35 in the next largest families (124 removed). The final dataset, PDB90t, had 1,937 sequences.

### *Calculating and Culling the Alignments*

Alignments were generated using gapped BLAST version 2.0.11 from the NCBI toolkit, downloaded from [ftp://ftp.ncbi.nlm.nih.gov/toolbox/ncbi\\_tools/ncbi.tar.gz](ftp://ftp.ncbi.nlm.nih.gov/toolbox/ncbi_tools/ncbi.tar.gz). BLAST was run with default alignment scoring parameters (BLOSUM62 scoring matrix, -11/-1 gap open/extend penalty). The BLAST “Expectation value” cutoff parameter for reporting alignments was set to 10,000, rather than the default value of 10.0, to force BLAST to report even very low similarity alignments. PDB90t sequences were aligned all-against all (1,072,242 alignments reported by BLAST). As this work is only interested in how *different* sequences can be compared, alignments of sequences to themselves were removed (1,937 removed). BLAST reports directional alignments, i.e. where swapping the query and database target (“sbjct”) roles of two sequences may affect their reported alignment and scores. Full SW algorithms may also be directional by reporting different

alignments in each direction when several optimal alignments exist at the same raw score. However, 20% of alignments returned by BLAST, an approximate SW, were reported in on—but not the other—direction. To avoid double-counting bi-directional alignments in my analysis, all alignments were made undirected by keeping the alignment of highest raw score and choosing randomly to break raw score ties (427,381 removed). The final alignment set, PDB90tU, had 642,924 alignments.

### *Measures and Formulas on Alignments*

To establish pairwise homology and non-homology between for PDB90t sequences without resorting to sequence-comparison methods, sequences were checked for having the same SCOP Class, Fold, and Superfamily; this YES/NO criterion was used to define Homologous/Not-Homologous alignments as recommended by (6).

Alignment length was calculated as the number of non-gap residue pairs in the alignment, while identities, positives, and expect score were taken as given by the BLAST. Visually, for BLAST output, it can be verified that

```
Query 1 CAGT--KGCKYFSDDGTFVCEG 20
      C G      G  + +  G      +G
Sbjct 6 CMGRGDSGGSWITSAGQ--AQG 26
```

would have alignment length 20, 5 identities, and 8 positives.

As noted in the Introduction, the percent identity of an alignment = % alignment identity = identities/(alignment length); the percent identity of two sequences = % sequence identity = identities/(shortest sequence length), where shortest sequence length = min(query sequence length, database target sequence length) = the maximum possible length of an alignment.



### *Analysis Tools and Miscellaneous*

All sequence files and BLAST output were parsed using custom Perl software and stored in an Oracle 8i™ relational database with SQL\*Loader™ (18). Statistics, tabulations, and curve fits were performed using the Statistica™ and R data analysis packages (23, 32), and custom software in Perl and Oracle 8i™ PL/SQL (18).

## **RESULTS**

### *Weakly Similar Alignment Set Compared to Twilight Zone and Non-Homolog Alignments*

The twilight zone cutoff included homologs with sufficient sequence similarity to warrant annotation as biologically similar based on SW alignment EVS scores, while excluding a significant number of non-homolog alignments. 1,797 non-homolog alignments were above the twilight zone cutoff, and 274 homolog alignments were below the twilight zone cutoff but had strong similarity (i.e. would be identified as biologically related based on their SW alignment similarities). All non-homolog alignments (100% = 633,610/633,610) were in the weakly similar set, and by definition no weakly similar homologs had clear sequence evidence for biological similarity. Nonetheless, the weakly similar set and twilight zone have substantial overlap. Over 95% of homolog alignments (5,736/6,010) and 100% of non-homolog alignments (631,813/631,813) that fell below the twilight zone cutoff were also weakly similar. 91% of homolog alignments (5,736/6,332) and over 99% of non-homolog alignments (631,813/633,610) that were weakly similar fell below the twilight zone cutoff. Overall, over 99% of weakly similar alignments were below the twilight zone cutoff (637,549/639,942), and vice versa (637,549/637,823).

#### *No. of Identities a Linear Function of Alignment Length for Distant Homologs*

Ninety percent of the variance in the number of identities in weak similarity alignments between homologs could be explained by the linear length-identity relationship, as shown in Figure 1 (a), 'Weak'. This relationship significantly weaker for strong similarity homolog alignments (Figure 1 (a), 'Strong'), perhaps explaining why the weak-homolog linear relationship has not been previously reported in empirical analyses which did not partition alignments by EVS scores' significance for separate analysis (6, 22, 24). The number of alignment residue-pairs with a positive BLOSUM62 similarity score ('positives', or 'P') was also linearly related to the alignment length ( $P=2.8 + 0.43*L + \epsilon(0,2.8)$ ,  $R^2=96\%$ ) for weakly similar homologs and, to a lesser degree, for close homologs ( $P=3.9 + 0.60*L + \epsilon(0,21.1)$ ,  $R^2=83\%$ ).

#### *No. of Identities a Linear Function of Alignment Length for Non-Homologs*

Non-homolog alignments demonstrated the same functional relationship, between alignment length and identities, as weakly similar homolog alignments (Figure 1 (b)). The number of positives and alignment length were also linearly related for non-homolog alignments ( $P=2.5 + 0.41*L + \epsilon(0,1.8)$ ,  $R^2=91\%$ ).

#### *Alignment Length Predicts Percent Alignment Identity and the HSSP Curve*

The relationship of alignment length and identity has been reported as a boundary, between homolog and non-homolog alignments, of percent alignment identity and alignment length (2, 22, 24). As Figure 2 empirically confirmed, the basic form of the widely used HSSP exponential was predicted by the linear relationship between alignment length and identity reported here. An HSSP-like boundary was reproduced by an upper confidence interval of the simpler linear relationship between alignment length

and identity (Figure 2, 'Boundary'). This is consistent with the conclusion of (2) that confidence intervals from a relationship between alignment length and identity reproduce the HSSP curve's shape. Other similarities between the results of (2) and those reported here appeared to be superficial. The theoretically obtained identity-to-length ratio of 5.8% given in (2) differs substantially from the 23%-25% slope observed here, and it was not evident that their average-residue-score arguments for calculating the identity over length  $I/L$  ratio, lack of a constant intercept term in relating  $I$  and  $L$ , and inclusion of a  $\sqrt{L}$  error term for predicting percent alignment identity were supported by my data. (E.g. a least squares fit for  $m$  of the equation  $I = m * L$  (2) to the same non-homolog data used in Figure 1 (b) gave  $m = 0.31$ ,  $R^2 = 65\%$ . Fitting  $I = m * L + Z' / \sqrt{L}$  gave  $m = 0.13$ ,  $Z' = 1.0$ ,  $R^2 = 80\%$ ; visual inspection of this curve (not shown) revealed close agreement with the linear relationship of Figure 1 (b), 'Weak', from  $0 < L < 100$  and poor fit to the data for  $L > 100$ .) Hence, to my knowledge, the linear relationship of alignment length and identity reported here, plus the empirical prediction of the HSSP boundary from this relationship, have not been previously observed.

## DISCUSSION

### *Linear Relationship Expected to Hold for SW Algorithms in General*

Approximate SW programs (e.g. BLAST (5), FASTA (21)) and full SW programs (e.g. MaxHom (27), SSEARCH (21)) implement approximations of the same theoretical algorithm (31) and are widely held to produce results that are essentially equivalent (3). Moreover, the alignment scatterplot density (Figure 2) and upper boundary (Figure 2, 'Boundary') relating percent alignment identity and alignment length are consistent with those reported for non-homologous alignments in previous studies (6, 22, 24). The spread

of data points above the HSSP Curve (Figure 2) was also similar to that observed in previous studies of comparable size (i.e. large) datasets (6, 22). These previous reports (6, 22, 24) used a mixture of different SW packages, data sets, homology criterion, scoring matrices and parameters. Hence, the a priori expected similarity for alignment characteristics from full- and approximate- SW programs, plus the observed agreement between previous studies and the data reported here, suggest that the functional relationship of alignment scores, identity, and length observed here hold for SW algorithms in general.

As data points can overlap in scatterplots, the precise quantitative distribution of data from previous studies cannot be determined from those reports (6, 22, 24). It cannot be ruled out that the particular constants (e.g. slopes, intercepts, and standard deviations in Figure 1) observed here would vary significantly under different experimental conditions. Nonetheless, the 23%-25% limiting ratio for percent alignment identity of weakly similar alignments predicted by a linear relationship in identity and alignment length (Figure 1) closely agreed with the 24.8% limiting HSSP value (24).

The relationships reported here may hold for alignment algorithms which (like SW methods) either act by extending alignments at their edges or evaluate alignments using scores derived from Sellers 1974 mutation + indel similarity metric. This may be tested by repeating the analyses of Figure 1 on alignments from structural alignment-by-extension methods (e.g. (7, 14, 30)) and generalized- or non-SW alignment tools such as profile alignment tools (e.g. PSI-BLAST (5)) or Hidden Markov models (c.f. (12)).

### *Basis of the HSSP Relationship*

The HSSP curve defining the boundary of percent alignment identity and alignment length for alignments between sequences of known and unknown homology is “a principal result” from the seminal work of (24). Those authors suggest that the HSSP curve exists due to a “physical theory of sequence-structure relation” (which they noted was unknown at the time of that report) (24). My results indicated that the percent alignment identity and alignment length boundary curve was a mathematical consequence of a strong linear relationship between alignment length and identity for weakly similar sequences (Fig 2, ‘Boundary’). This linear relationship held for non-homolog as well as homolog alignments (Figure 1), indicating that structural similarity between the sequences aligned was not required to explain an HSSP-like boundary. Rather, HSSP-type boundaries were observed to be artifacts of sequence-alignment algorithms and scoring systems that generated a linear relationship between alignment length and identities for weakly similar sequences.

### *Percent Identity and Similarity Measures Deprecated*

Ratios of sequence length to identity or positives based on SW alignments are widely used measures of similarity (e.g. (6, 10, 22, 24, 26, 34)). I suggest these be deprecated for weakly similar sequences. Given that weakly similar alignments’ length strongly predicted their number of identities, dividing one by the other appeared uninformative for evaluating biological- or sequence- similarity between weakly similar sequences. For shorter alignments, the wide range of percent alignment identity was explained by the random error term  $\varepsilon$  (c.f. Figure 1). As  $\varepsilon$  had little dependence on alignment length, dividing it by an increasingly short alignment length resulted in an increasing large

spread in percent alignment identity. For longer alignments, the percent alignment identity converged to the baseline of 23%-25%. Hence, graphs and analyses using percent alignment identity for SW alignments must be carefully interpreted with the understanding that differences in percent alignment identity are unlikely to indicate differences in sequence similarity or biological similarity between weakly similar sequences. This caution applies to most datasets capturing alignments for “distant” homologs (e.g. (6, 22, 24, 26, 34)), since I observed that over 99% of sequence pairs below the 25% sequence identity twilight zone cutoff were weakly similar.

Percent sequence similarity does not include an alignment length denominator, and appeared a valid measure of sequence similarity for weakly similar sequences. However, for SW alignments of weakly similar sequences I observed that  $(I/L_{max}) \approx 0.25 * L/L_{max}$  (c.f. Figure 1), indicating that percent sequence identity was a linear re-scaling of fraction aligned (e.g. as used by (26)). As a measure of sequence similarity for weakly similar SW alignments, fraction aligned was preferable to percent sequence identity because it captured the intuitive physical basis underlying percent sequence identity: the ratio of the observed alignment length and the maximum possible alignment length.

My analysis of percent identity measures relied on the linear relationship of a raw similarity score and alignment length, rather than the weight given to each sequence-pair match. Like identities, SW alignment positives was well predicted as a linear function of alignment length for weakly similar sequences. Hence, “percent similarity” ratios of positives (or other properties linearly related to alignment length) to alignment length were deemed uninformative-at-best as measures of sequence similarity for SW alignments of weakly similar sequences.

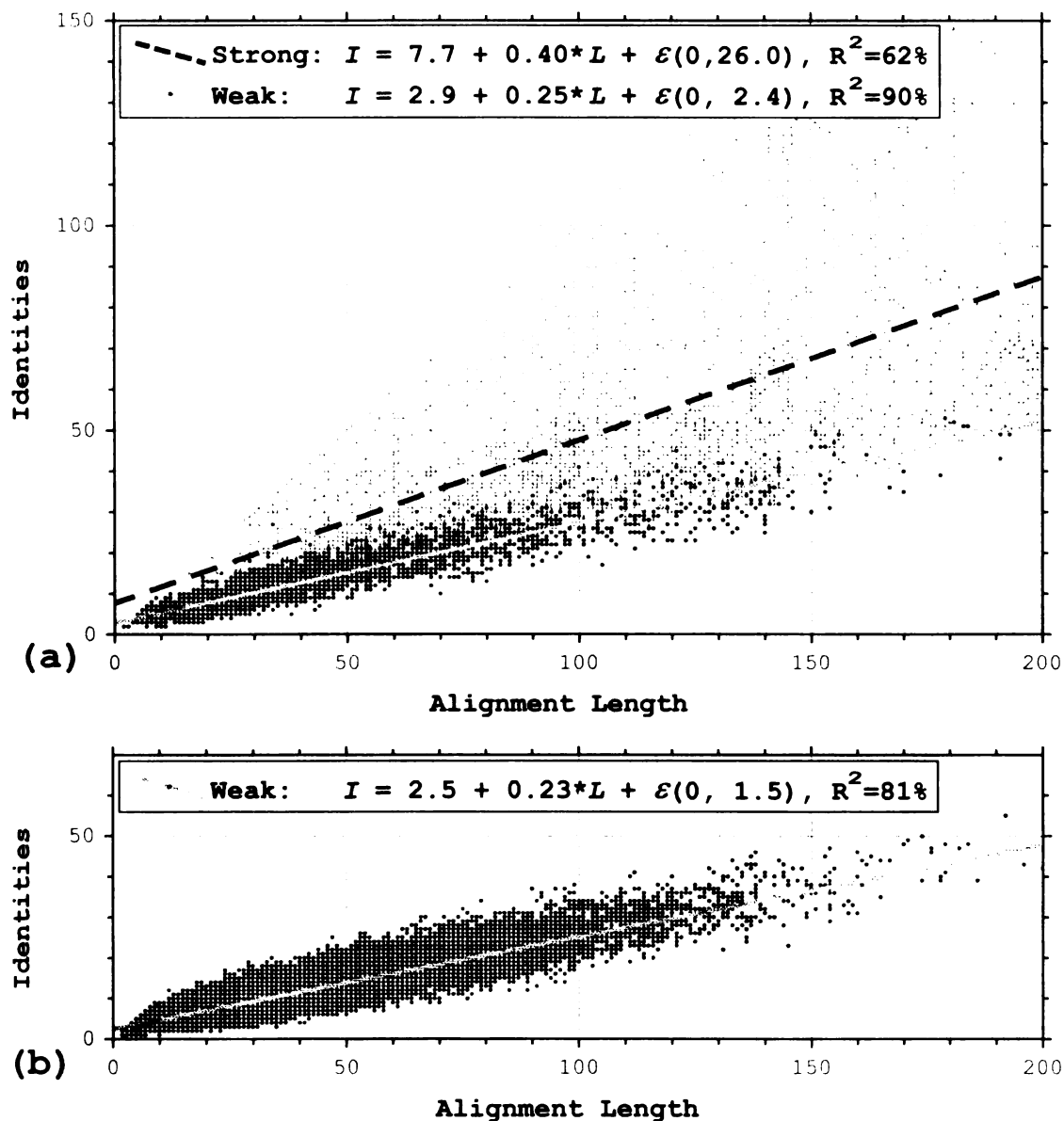
## CONCLUSION

It is widely recognized that the availability of reliable measures of sequence- and biological- similarity for alignments between distantly related genes is a principal barrier to understanding the biomedical properties of genes and genomes (c.f. (6, 12, 22, 26)). By analyzing sequences sharing sequence similarity of low statistical significance, I identified a strong empirical linear relationship between the length of alignments between those sequences and the number of identical (and similar) residues reported in their alignments. For these sequences, the alignment length predicted the observed number of identical (and similar) residues. Widely used ratio measures of sequence similarity, such as percent alignment identity, thus reflected virtually no information about actual biological- or sequence- similarity for weakly similar or distantly related sequences. The currently accepted relationship between alignment length and identity, the HSSP boundary noted as a principal result of (24), was extended by demonstrating the HSSP boundary could be reproduced by using alignment length alone to predict percent alignment identity. Because they address basic measurable properties of any sequence alignment, the experiments reported here can be repeated on the full range of methods for comparative sequence analysis, including local structural alignments (e.g. (7, 14, 30)), PSI-BLAST (5), and Hidden Markov Models (c.f. (12)). Such an analysis may reveal whether the apparently non-biological length-identity relationship is a limitation of current algorithms for calculating sequence alignments or the measures of similarity used to score them.

## ACKNOWLEDGEMENTS

I thank Dr. Mark R. Segal for insightful suggestions and generous help in statistical analysis for this work, Dr. C. Anthony Hunt for patient guidance and discussions, and DBA Bryan W. Taylor for contributing many miracles in Oracle database administration and SQL tuning. R.P.O. was partially supported by NIH grant GM08388; additional computing hardware was purchased through an award from the Regents of the University of California.

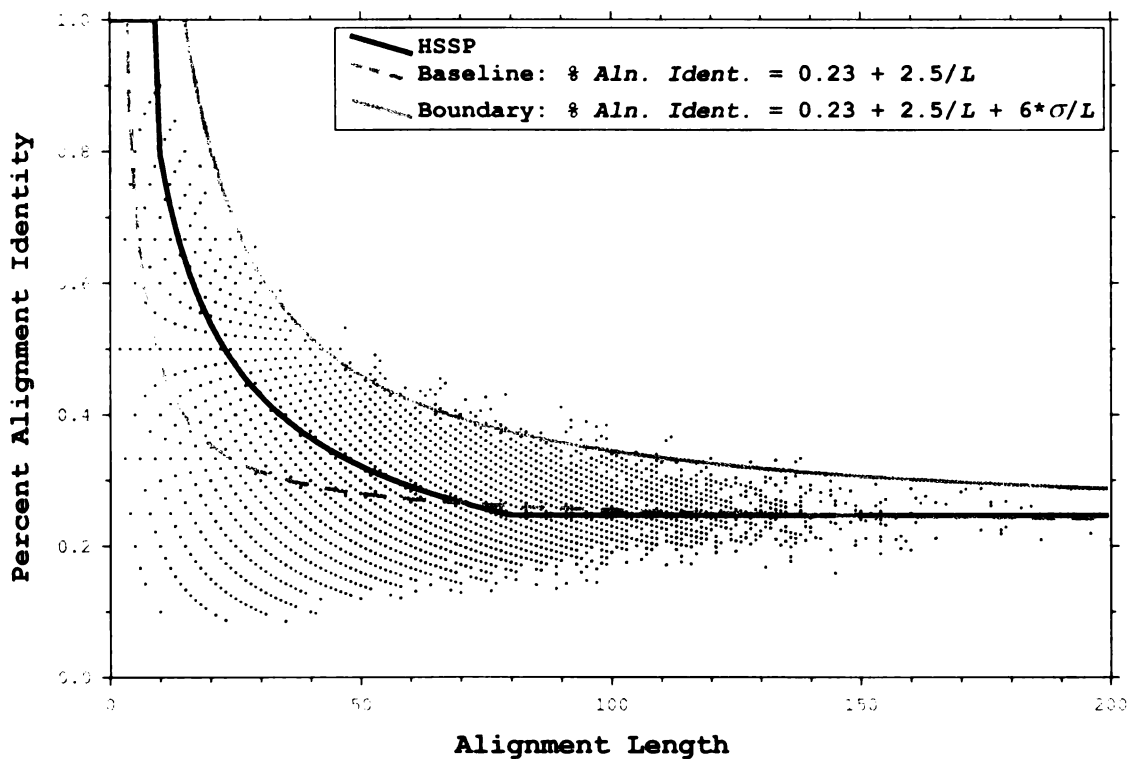




**Figure 1.** Alignment length versus number of identities for all PDB90tU alignments. For each alignment subset, the graph shows a least squares linear fit with error term  $\varepsilon(\mu, \sigma)$ , where  $\mu$  and  $\sigma$  are the mean and standard deviation of the fit error term, and  $R^2$ , the percentage of the variation in the alignment subset explained by the given line and equation. For simplicity, a  $10^{-5}$  cutoff for strong/weak alignment similarity was selected as comparable to a 1% marginal significance cutoff for directed alignments from 1,000 sequences using standard directed BLAST alignments. Note that these curve fits model

the **emp**irically observed alignment data, and are not intended to address theoretical alignments of arbitrarily short length (all 642,924 PDB90tU alignments had length > 10).

**(a) All** homolog alignments. Light triangles ('Strong' equation) are the homolog pairs with a BLAST expect score less than  $10^{-5}$  (6,332 alignments); dark circles ('Weak' equation) are the homologs at or greater than  $10^{-5}$  (2,976 alignments). **(b)** All non-homolog alignments. All non-homolog alignments in PDB90tU had a BLAST expect score greater than  $10^{-5}$ , and are shown as dark circles ('Weak' equation) (633,610 alignments).



**Figure 2.** Alignment length versus percent alignment identity for all non-homolog PDB90tU alignments. Dark circles show the values of alignment length and percent alignment identity as reported by BLAST. The ‘Baseline’ shows the mean relationship between percent alignment identity and alignment length predicted by a linear relationship between alignment length and identities for non-homologs. The ‘Baseline’ formula was calculated as  $I^*/L$ , where  $I^*$  equaled  $I$  as predicted from alignment length using the ‘Weak’ equation of Figure 1 (b) with the error term  $\varepsilon$  set to zero. By retaining a non-zero error term, the ‘Boundary’ curve illustrates the functional agreement between confidence intervals predicted by Figure 1 (b), ‘Weak’, and the empirically observed upper boundary of non-homolog alignment length and percent alignment identity values. The interval was chosen for visual clarity, as all but 128 of the 633,310 non-homolog alignments fell within six  $\sigma$  ( $\sigma = 1.50$ ) of the identities prediction line in Figure 1 (b), ‘Weak’. The original ‘HSSP’ boundary curve, calculated using different parameters and

smaller datasets (24), is included for comparison. The 'HSSP' formula is percent alignment identity = {1.00 if  $L < 10$ ;  $2.9015 * L^{-0.562}$  if  $10 < L < 80$ ; 0.247 if  $L \geq 80$ } (24).

The elegant arcs and swirls observed in the plotted points appeared to be Moiré patterns (caused by plotting integers and their ratios) with no biological significance. Note that each plotted point may represent many alignments sharing the same coordinates.

## REFERENCES

1. **Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, *et al.*** 2000. The genome sequence of *Drosophila melanogaster*. *Science*. **287**(5461):2185-95.
2. **Alexandrov, N. N., and V. V. Solovyev.** 1998. Statistical significance of ungapped sequence alignments. *Pac Symp Biocomput*:463-72.
3. **Altschul, S. F., M. S. Boguski, W. Gish, and J. C. Wootton.** 1994. Issues in searching molecular sequence databases. *Nat Genet*. **6**(2):119-29.
4. **Altschul, S. F., and W. Gish.** 1996. Local alignment statistics. *Methods Enzymol*. **266**:460-80.
5. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**(17):3389-402.
6. **Brenner, S. E., C. Chothia, and T. J. Hubbard.** 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A*. **95**(11):6073-8.
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.
8. **Consortium, T. C. e. S.** 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*. **282**(5396):2012-8.
9. **Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt.** 1978. A Model of Evolutionary Change in Proteins. *Atlas of Protein Sequence and Structure*:345-352.

10. **Doolittle, R. F.** 1986. Of urfs and orfs : a primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley, CA.
11. **Doolittle, R. F., M. W. Hunkapiller, L. E. Hood, S. G. Devare, K. C. Robbins, S. A. Aaronson, and H. N. Antoniades.** 1983. Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science*. **221**(4607):275-7.
12. **Durbin, R., S. Eddy, A. Krogh, and G. Mitchison.** 1998. Biological sequence analysis : probabalistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, UK New York.
13. **Henikoff, S., and J. G. Henikoff.** 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*. **89**(22):10915-9.
14. **Holm, L., and C. Sander.** 1995. Dali: a network tool for protein structure comparison. *Trends Biochem Sci*. **20**(11):478-80.
15. **Karlin, S., and S. F. Altschul.** 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*. **87**(6):2264-8.
16. **Madden, T.** 2000. NCBI Software Development Toolbox, minimum query length calculation, file://tools/blast.c, Mon Oct 23 20:20:58 PDT 2000 ed. The toolkit is available by anonymous ftp from ftp://ncbi.nlm.nih.gov.
17. **Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia.** 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. **247**(4):536-40.

18. **Oracle.** 2000. Oracle8i Enterprise Server, Release 8.1.5 ed. Oracle Corporation, 500 Oracle Parkway, Redwood City, CA 94065.
19. **Page, R. D. M., and E. C. Holmes.** 1998. Molecular evolution : a phylogenetic approach. Blackwell Science, Oxford ; Malden, MA.
20. **Pearson, W. R.** 1998. Empirical statistical estimates for sequence similarity searches. *J Mol Biol.* **276(1):**71-84.
21. **Pearson, W. R.** 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol.* **132:**185-219.
22. **Rost, B.** 1999. Twilight zone of protein sequence alignments. *Protein Eng.* **12(2):**85-94.
23. **R-Project, D. C. T.** 2000. R - A language and environment for statistical computing and graphics. The R Project for Statistical Computing.
24. **Sander, C., and R. Schneider.** 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins.* **9(1):**56-68.
25. **Sander, C., and R. Schneider.** 1993. The HSSP data base of protein structure-sequence alignments. *Nucleic Acids Res.* **21(13):**3105-9.
26. **Sauder, J. M., J. W. Arthur, and R. L. Dunbrack, Jr.** 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins.* **40(1):**6-22.
27. **Schneider, R.** 1994. PhD, University of Heidelberg. .
28. **Schuler, G. D., M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tome, A. Aggarwal, E. Bajorek, et al.** 1996. A gene map of the human genome. *Science.* **274(5287):**540-6.

29. **Sellers, P.** 1974. On the Theory and Computation of Evolutionary Distances. *J. Appl. Math (Siam)*. **26(4):787-793.**
30. **Shindyalov, I. N., and P. E. Bourne.** 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11(9):739-47.**
31. **Smith, T. F., and M. S. Waterman.** 1981. Identification of common molecular subsequences. *J Mol Biol.* **147(1):195-7.**
32. **Statsoft.** 1999. Statistica, 5.5 ed. Statsoft, Inc., 2300 East 14th Street, Tulsa, OK 74104.
33. **Waterman, M., T. Smith, and W. Beyer.** 1976. Some Biological Sequence Metrics. *Advan. Math.* **20:367-387.**
34. **Wood, T. C., and W. R. Pearson.** 1999. Evolution of protein sequences and structures. *J Mol Biol.* **291(4):977-95.**



## ENTERING BIOINFORMATICS

I began bioinformatics research as a side-project, by looking for an interesting problem in computational biology that matched my toolbox. To wit, I had begun research in an experimental lab and sought an outside project to keep my skills in mathematics and computer science fresh and sharp.

Earlier work in theoretical protein folding and molecular docking simulations drove me to seek research where an answer could be easily tested. This is currently not practical for the sub-nanosecond motions in a macromolecular-folding ensemble or the predicted rank-order of interaction energies for compounds dissolved in a time-averaged protein structure. Thanks to genome projects and mutational studies, orthogonal data is often available to evaluate a proposed gene family's evolutionary tree or conserved catalytic motifs, thus making validated bioinformatics a fairly practical sport.

My passion for bioinformatics continues to grow as I see what is possible, but left undone. Biology is beginning a transition much like early 20th century physics, when an accumulation of data shattered the Newtonian view and allowed relativity and quantum theory to be discovered. Starting in the 1990s, high-throughput experimental biology began to provide the data required to predictively and quantitatively understand the mechanical parts, regulatory interactions, and information processing machinery of living systems. Medicine, science, and philosophy will make tremendous advances based on subtle patterns extracted from accumulating petabytes of measurements on living systems. I hope to contribute to that.

Before I'd even thought about entering bioinformatics I had worked very hard to understand mathematics, and then biology, at the graduate level. So it's fair to say I was an interdisciplinary person who entered the field, rather someone who prepared for it by name. My principal preparation has been to follow my interests, striving to master the hardest course in a new discipline shortly after entering it, and committing the time and struggle required to become a peer with top students on their own turf in each discipline.

### **MY PREPARATION FOR RESEARCH**

My formal preparation started at home with a precious new 48K Apple ][+ computer and, later, biology in 10th grade. On entering college at the University of Texas, Austin, I had to choose between biology and computer science (CS) majors. An underinformed vision of personally collecting measurements from ten thousand neatly stacked petri dishes drove me to computing. My background allowed me to test out of some lower-division courses and take the CS core concurrently, which gave me the choice to graduate in two years. But my freshman CS coursework was more application than theory, and left me feeling more knowledgeable but like my mind hadn't been stretched far enough. So I decided to take a four-year plan where I explored physics, philosophy, and mathematics. After my sophomore year I speculated that mathematics was somehow the natural language of nature, and took a math class pretty much every semester, including summers, until I graduated with a BA in mathematics and a BS in computer science. Ultimately, my college work gave me helpful preparation in applied computer science, artificial intelligence, logic, linguistics, wave mechanics, electrodynamics, quantum mechanics, fractals, differential equations, vector-tensor analysis, geometric linear algebra, and knot theory. At the time I regretted missing organic chemistry and intensive

essay writing, but fortunately managed to pick these up later. (I'm still working on graph theory and advanced algorithm design.)

For largely romantic reasons, I enrolled in the mathematics Ph.D. program at UT Austin and was immediately slammed into the local bone-crusher course, Real Analysis with Professor William Beckner. Every word of these brilliant lectures baffled my office-mates and me as we labored in our windowless office, trying to conjure intuition about 'getting some control . . . over this set of duals for infinite dimensional spaces of smooth functionals with compact support' (a.k.a. the graduate view on floating point numbers). (Of note, graduate and undergraduate analysis are basically unrelated, particularly if you haven't had point-set topology.<sup>25</sup>)

This seemingly futile struggle consumed four months of dawn-till-dusk struggle without a glimpse of true understanding. It was only during my second year, as I caught myself saying '. . . well, we have a little control under here . . . so if you can squeeze with that inequality, you'll get convergence' that I realized that this course—which remains my most treasured educational experience—had hammered down a lasting foundation for rigorous axiomatic problem-solving.

A simultaneous and unexpected preparatory for bioinformatics was my mother's sudden illness and death from lung cancer during my second semester. This, coupled with a physicist's good advice to "think really hard about what your whole career's twenty or thirty papers will do for this world", created an enduring energy and resolve for focusing

---

<sup>25</sup> In retrospect, I highly recommend working through the standard *Topology: A First Course* by JR Munkres (Prentice Hall, 1974) and then the less-well-known *Real Analysis: Modern Techniques and Applications* by GB Folland (John Wiley and Sons, 1984) as a hard but streamlined route to thinking in a structured, mathematical way.

my time towards building widely useful tools for preventing disease and intensive-care-unit suffering. I mention this because my path, forging across several hard fields towards an unseen goal, brought many periods of professional doubt and some narrow escapes from the long arm of a few entrenched traditionalists. Instantly quadrupling my income with a lucrative career in programmer-starved Silicon Valley has remained a tempting option since my freshman year in college, as has choosing an easier class or accepting a standard-but-poorly-understood experiment. So tremendous resolve, aided by encouragement from friends and by luck, has proven essential for pursuing science.

My post-highschool training in biology began with the transition away from pure mathematics. Not knowing the difference between microbiology and biochemistry, I visited a visionary biologist who once gave me pocket money in exchange for washing his lab's glassware. He recommended protein folding and mining genomic data as the top challenges in biology suited to my background. On his suggestion I also took physical chemistry, graduate biochemistry, and two semesters of organic chemistry while applying to biological Ph.D. programs. Like many physics and math students entering biology, I chose to work on understanding biological systems from physical first principles, and hence applied to the UC San Francisco Biophysics group to train under a well-known giant in protein folding theory, Professor Ken Dill. While focusing primarily on computational lab-work, I struggled to get a grip on this new field through the standard graduate courses in spectroscopy, statistical mechanics, structural biology, drug-design, macromolecular interactions, and experimental techniques in biophysics. After these studies I felt I still couldn't think like a true experimental biologist, so with this kind advisor's blessing I labored through the Biochemistry program's core graduate studies in

genetics, cellular biology, cell signaling, and biological regulatory mechanisms. This last course, BioReg, was widely held as the hardest course at UCSF, with a fail rate rumored at 25%-40%. The course, with a dreaded oral cross-examination final, was based on lectures coupled to critical discussions of the scientific literature. Two other students and I supplemented these with weekly meetings for exchanging impromptu challenge-and-answer exercises like ‘In class Joachim said that the DNA replication fork clamp is processive; how would you experimentally demonstrate that if given the purified proteins and one week in a standard UCSF lab?’ Like Real Analysis at Texas, many courses had provided knowledge related to my current studies. But none truly forced me to use all available angles on the subject to try and work through hard questions, many times a week, in camaraderie with well-trained students whose research careers would be in that field. BioReg forced me to adopt the thinking skills that have proven effective to biochemists and geneticists for zeroing in on truth in a complex interacting system. Trying to identify a minimal basis of facts and propose specific answers consistent with that basis, as I’d learned in mathematics and physics, simply made poor use of the available data without producing useful predictions. Of note, a jewel among these techniques was to develop experience in proposing controls to prevent reasonable but unjustifiably specific conclusions, while devising assays to distinguish which broad answers encompass a far narrower physical truth. The intersection of broad answers from several assays, safeguarded by the right controls, could usually be applied to quickly close in on a useful version of the truth. This also had the side-benefit of suggesting additional experiments and testable predictions to further sharpen the picture.

The rest of my training—including a year in experimental high-throughput assays and my present research in computational bioinformatics with Professor Tony Hunt—has largely been founded on self-study guided by many conversations with other scientists and kind faculty mentors.

#### **WHAT I'VE FOUND TO BE USEFUL**

*For scientific training:* Striving to practice thinking skills, particularly through critical reading and working through hard problems in camaraderie with peers who will be researchers the topic's field. Attending a conference early on to get a sense of the field.

*For guidance and collaboration in an uncommonly cross-disciplinary field:* Building a network of supportive peers and mentors, based on people with a track-record of good character rather than famous science. Building mutually supportive interactions rather than accepting a competitive academic mindset. Seeking advice from multiple sources with different perspectives and backgrounds.

*For efficient productivity:* Finding a source of funds for a laptop with software for solitary thinking and computing in the library, and for purchasing unrequired, exploratory books. (I find note-taking in the margins and multi-colored highlighters are key thinking tools.)

*For thriving in academia:* Discovering that publications, rather than knowledge or ideas, are the primary currency of biological academia. Productivity evaluations, continued paychecks, and institutional acceptance are directly determined by your current or anticipated publication record—with education provided as a reward and a means to

generate better publications. Discarding the belief that faculty treat students as peers, rather than free-lance employees who *may* become peers *after* the Ph.D.

*For long-term success:* Striving for practical honesty and good character under pressure. Hard work and optimism.

#### **WHAT I HAVE FOUND NOT USEFUL**

*Pitfalls of Judgement:* Using recent history, rather than long-term track record, as a guide to predicting people's actions in a complex organization. Focusing on laboratory work rather than deep education during my first year. Jumping into a hot project rather than beginning with writing down a plan of action, milestones, anticipated risks, introductory background reading.

*Obstacles of Circumstance:* Lack of biologists' infrastructure-funded support for network bandwidth, computing horsepower, information technology, software purchasing, high-throughput data-generating equipment, and partnerships with data-rich industrial partners. Pay rate qualifying poverty-line subsidies for electricity—in my late twenties. Seeing others suffer due to years of underfunded healthcare in graduate school. Lack support for basic research; I need to eat and also to feel like I have a future *in my current field*, which is why I work so hard to learn the foundation that will carry me for twenty years, rather than five. Currently, a Ph.D. removed from immediate commercial impact appears guaranteed to be an exercise in preparing to work in another field if I want to have a house and family as well as to contribute to science.

## **FOR BUILDING SUSTAINABLE CAREER SKILLS**

I've developed a fair intuition about molecular evolution, and have a rack of my own computational assays to apply in this area, so right now I'm building the machinery to do terabyte-to-petabyte scale genome analysis. Very large databases, component-based computing, and statistical techniques for large data-sets are all lacking in bioinformatics toolkits right now. Building these tools will allow me to tackle the big questions of biology for years to come. Our group should also be able to broaden our collaborative base fairly soon by publishing some of this work for use by other people.

Strategically, I definitely see that there is a tremendous advantage in choosing to nurture the therapeutic applications of my work. Research aligned with a therapeutic application is better funded, draws in a broad group of brilliant colleagues, and has the personally energizing potential to accelerate the cure of a major disease by 5-10 years. I'm still doing basic research, but fortunately there happen to be substantial applications for aspects of my work. Since I generally have at least 10-20 interesting ideas to pursue at a given moment, I tend to choose the ones that will have biomedical applications even if it adds extra overhead to develop them.

A point of note is that some strengths, like statistics, computing, or deep biological intuition, take years to build and are also broadly useful. Anyone can pick up a quick technique, but some abilities require multiple layers of training which are tractable at each step but, collectively, are guaranteed to require years to attain<sup>26</sup>. That means other

---

<sup>26</sup> Having taught many students as a tutor and teaching assistant, I believe this is why many have trouble with physics and math. Each course is tractable, but only if you have all the layers below. Taking electromagnetics before geometric calculus, or geometric



scientists needing those skills will have to spend those same years laying down those skills, which gives me a modicum of security in a field of rapidly changing technologies. You can bet I quietly add another layer to some long-ramp-up-time skill on a regular basis.

#### **ACTIONS THAT PREPARE FOR UNANTICIPATED CHALLENGES**

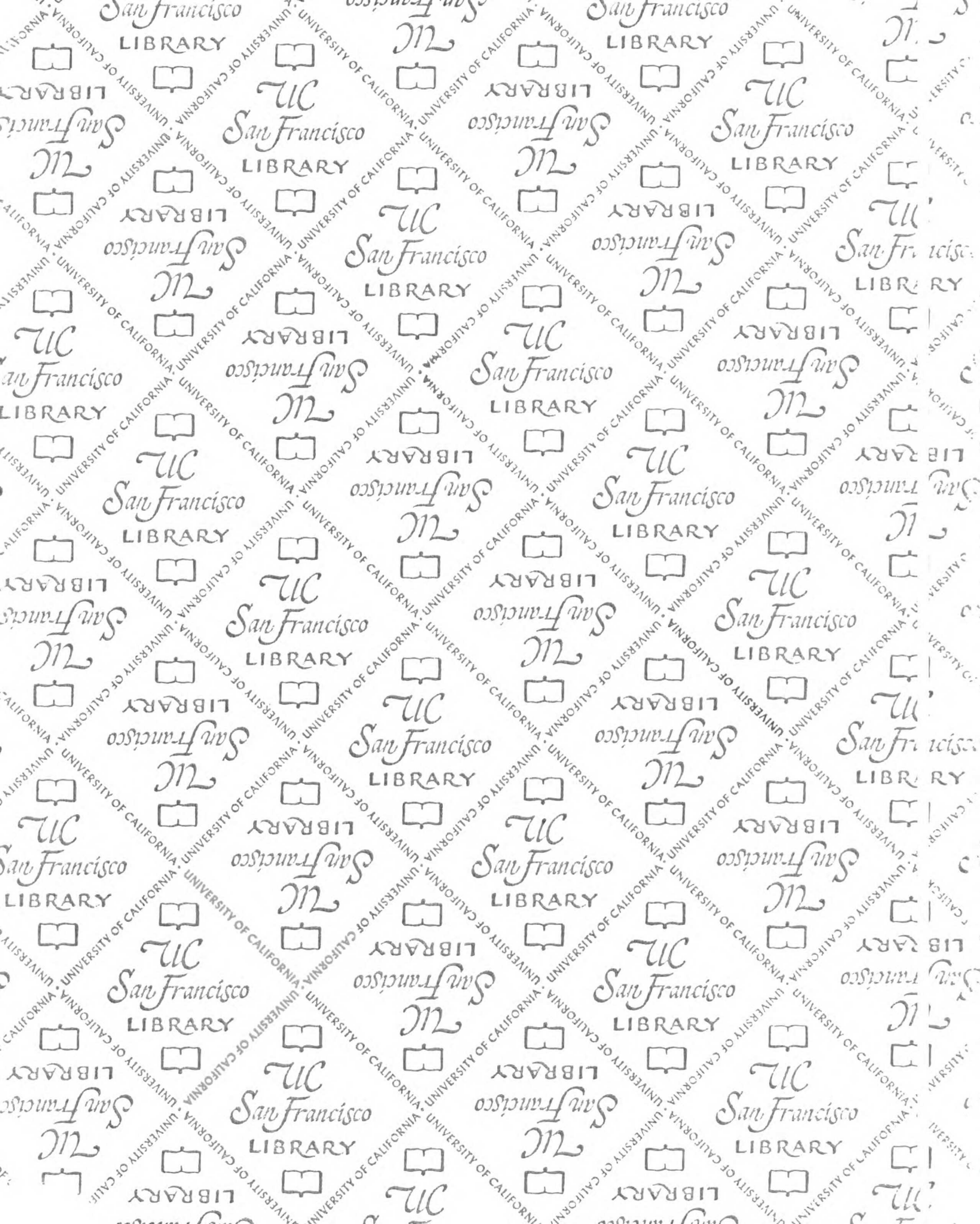
1. Building long-term relationships with people of good character, and seeking their advice. I have been saved countless times by good unsolicited advice from kind people, so I try to seek it more actively when I can.
2. Looking for synergistic collaborations with people who enjoy doing things I don't, and vice-versa, so that we form a fun, synergistic team in working towards a common goal. I find these types of partnerships are more robust to challenges, thanks to a broadened skill-base and many informed perspectives to anticipate challenges from sides I that wouldn't be trained to recognize.
3. Checking trade journals for technology that will provide data or techniques in my field
4. Watching funding trends for new sustainable directions in the field and keeping my Silicon Valley skills up-to-date. These, and my long-ramp-up-time strengths, could pay the bills and keep me in science if bioinformatics crashes.
5. Trying to follow broad trends in politics, economics, technology, and society.

---

calculus before algebra, is like being forced to understand poetry in Spanish before you've learned basic Spanish grammar: hard and frustrating.

6. Taking time to help people who are already working hard to help themselves without stepping on anyone else; this has had a big return in the past and is an easy way of giving back despite my tight schedule.
7. Reminding myself that I'm helping an adventure of discovery that will change our understanding of nature as fundamentally as the Newtonian/Euclidean→Relativistic/Quantum view did in the early 1900s. Reasonable facts indicate this will happen during my career, which encourages me to hang on during the rough spots.

*Author's Note of Dedication: My deepest appreciation to John Spikerman, William Beckner, Ken Dill, Tony Hunt, Julie Ransom, and many, many kind people who selflessly encourage and support my growth and research. My ability to follow my dream is the fruit of their patient advice, timely support, and guided license to pursue my own path.*



# For reference

Not to be taken from the room.

7063533



3 1378 00706 3533

