

INVESTIGATING EVOLUTIONARY LINES OF LEAST RESISTANCE USING THE INVERSE PROTEIN-FOLDING PROBLEM

J. Schonfeld, O. Eulenstein, K. Vander Velden, and G. J. P. Naylor
Bioinformatics and Computational Biology
Iowa State University
Ames, IA 50011

We present a polynomial time algorithm for estimating optimal HP sequences that fold to a specified target protein conformation based on Sun et al's Grand Canonical (GC) model. Application of the algorithm to related proteins taken from the PDB allows us to explore the nature of the protein genotype:phenotype map. Results suggest: (1) that the GC model captures important biological aspects of the mapping between protein sequences and their corresponding structures, and (2) the set of sequences that map to a target structure with optimal energy is affected by minor differences in structure.

1 Introduction

1.1 Background

A fundamental problem in biology is to understand the correspondence between the genotype and phenotype. Understanding the functions that constitute the *genotype:phenotype map* (*gp-map*), would facilitate the prediction of genetic predisposition to disease, the design of new drugs, and an understanding of the evolutionary origin of the diversity of phenotypes. The complexity of biological systems has made it extremely difficult to elucidate the *gp-map* at the organismal level. Our current knowledge of this map is generally limited to a few alleles, most of which are associated with human disease conditions. Recent efforts by the biophysics community to understand the *gp-map* at the level of proteins look particularly promising [1,2], and may provide insight into higher levels of organization.

Proteins represent an excellent system in which to explore the *gp-map* for two reasons:

- 1 They exhibit a broad range of functions (phenotypes) including cell signaling, pathogen recognition, structural support, cellular scaffolding, and molecular motors that move components around within the cell.
- 2 There is a large body of work devoted to solving the *protein folding* problem, which is defined as follows: Given a protein sequence S , find the conformation to which S folds under physiological conditions. Research in this area has led to an improved understanding of the rules that govern how sequences map to their corresponding protein phenotypes.

1.2 Redundancy and Accessibility

Studies of protein variation in nature indicate that there is extensive redundancy in the mapping from sequence to structure [3,4]. It is likely that the highly redundant mapping between sequence and structure has been important for the evolution and diversification of proteins [5,6,7]. Structures with many sequences mapping to them are predisposed to be more accessible through the evolutionary process than are structures with fewer sequences. For example, consider two sequences *A* and *B*, each of which maps to a different structure, and which differ by 20 point mutations. Changing one structure into the other would require the simultaneous mutation at each of the 20 different sites. This is highly unlikely to occur. However, if the same two structures are each represented by several thousand sequences, there is a higher likelihood that some of the sequences mapping to *A* would be closer to some of the sequences mapping to *B*. Thus a high degree of redundancy in the mapping between *A* and *B* promotes their mutual evolutionary accessibility.

A protein can be thought of as a *cloud* of points in a high dimensional space, where each axis represents a separate amino acid position in the sequence. This space is non-Euclidean and is referred to as *protein space* [6,8]. Each point in the protein space represents a unique sequence while a cloud represents the domain of different sequences that map to a particular protein's function. When clouds representing different proteins come into close proximity, a change from one protein to another is facilitated. Change is unlikely when clouds are distant. Just as spherical clouds will lead to less connectivity than will dendritic clouds in Euclidean space, the multidimensional shape of protein clouds will affect the degree of connectivity and accessibility among proteins. Comparable dendritic clouds have been investigated as sparse random networks [9] and as neutral networks [4,10].

In this paper we set out to use the redundancy in the mapping of sequence to structure to explore the question: "How easy is it to turn one kind of protein into another, and are there paths of least resistance that would allow us to best account for the diversity of proteins observed in nature?" More formally, what is the evolutionary inter-accessibility among proteins? We call this the *minimum evolutionary distance (MED) problem*. To solve the MED problem and examine the structure of these protein clouds we need to have access to a large percentage of all of the sequence variants that occur in nature.

1.3 Previous Results

It is not yet possible to empirically collect all sequence variants that occur in nature to solve MED. Given this we generate sequence variants based on an abstract folding model. There is a natural inverse of the protein folding problem, called the *inverse protein folding problem (IPF)*, which has been used by many researchers to

tackle protein design problems [11]. For our purposes it is sufficient to describe the problem as follows: Given a native protein conformation C (*known as a target conformation*), find a sequence that folds to C with the minimal energy. For more detailed information the reader is referred to references [11,12,13]. A variety of methods have been described that attempt to solve IPF [14,15,16]. Their utility varies with their ability to capture different aspects of the problem and their computational complexity. Only a few of the approaches used are computationally solvable in polynomial time. Almost all are based on Dill’s HP-lattice model [17]. In Dill’s model a protein conformation is a self-avoiding walk on a regular lattice, and amino acids are classified as either hydrophilic (P), or hydrophobic (H). Two amino acids are assumed to be in contact if they are close in space, but not adjacent on the self-avoiding walk. The energy function rewards only H-H contacts. A considerable body of work suggests that this abstraction can capture important aspects of protein structure [12, 16, 17]. Following earlier models proposed by Shaknovich and Gutin [2], Sun et al. [1] introduced the Grand Canonical (GC) model. This model accommodates important aspects of the 3D conformation of real proteins by relaxing the regular lattice constraint and incorporating solvent accessibility. When applied to real proteins the GC model yields HP-sequences that closely match the HP representation of real sequences [1]. Kleinberg [18] showed that IPF for the GC model can be solved in polynomial time by transforming it into a bipartite network flow problem.

1.4 Contribution of this Paper

In this paper we introduce the MED problem. We present and study aspects of the MED problem as it applies to biological evolution and the “interaccessibility” of proteins. Our approach builds on solutions to IPF based on the GC model and uses network flow techniques initiated by the work of Kleinberg. We apply the mathematical structure and efficient algorithms associated with network flow problems to address issues related to the MED problem. We report the following:

Computational advances:

- An improvement in the running time of Kleinberg’s algorithm.
- An efficient representation and algorithm to find all HP-sequences that optimally solve IPF for the GC model.

Evolutionary advances:

- Comparable accuracy to Sun et al.’s model [1] when estimated HP-sequences are contrasted with the corresponding real sequences.
- A demonstration that minor changes in protein structure can have important consequences for the sequences that map to them with minimum energy.

- The surprising finding that the evolutionary information inherent in minor structural differences is reflected in HP-sequences that solve IPF for the GC model.

2. Solving the Inverse Folding Problem in “Computer Space”

The GC model as proposed by Sun et al. [1] is a HP-model that abstracts physico-chemical and geometrical features of real proteins into a contact graph G . Amino acids of the protein correspond to the nodes of G , and distances between amino acids below a given threshold correspond to edges of G . Edges are weighted by distance, and nodes are weighted by the solvent accessibility. The contact graph represents the target conformation. Hs of a HP-sequence map to a subset of the nodes of the contact graph and are referred to as an H-assignment. An H-assignment has an energy value given by an energy function that balances the competing cost of solvent accessibility and H-H contacts. H-H contacts are rewarded proportional to their distance, while H assignments are penalized proportional to their solvent accessibility.

In this section we provide a definition of the GC model that we use as the basis of our work. We then define IPF under the GC model and refer to it as GC-IPF. HP-sequences that solve GC-IPF are referred to as *optimal HP-sequences*. The definitions we give follow Sun et al. [1] and are presented in subsection 2.1. In subsection 2.2 we review the conceptual basis of Kleinberg’s algorithm to solve IPF under the GC model, and introduce an algorithm that is asymptotically more efficient. In subsection 2.3 we show that the work of Picard et al. [19] can be applied to reveal the intrinsic structure of all IPF solutions. This intrinsic structure is also used to enumerate all IPF solutions. For a background in standard network flow concepts the reader is referred to [20].

2.1 Grand Canonical (GC) Model

Definition (contact graph): Let C be a protein structure. A contact graph $G_c = (V, E, s, d)$ is a simple undirected graph, with vertices V , edges E , node weights $s: V \rightarrow R$, and edge weights $d: E \rightarrow R$. Vertices in V correspond to the amino acids of the conformation. Edges in E correspond to non-covalent binding amino acids, whose C_α positions are at most 6.5\AA apart. The node weight $s(v)$ represents the “solvent accessibility” of the corresponding amino acid v and the edge weight $d(\{v,w\})$ represents the “distance” between the amino acid residues corresponding to v and w . (In subsection 3.1 we will specify “solvent accessibility” and “distance”.) An example for a Contact Graph is shown in Figure 1.

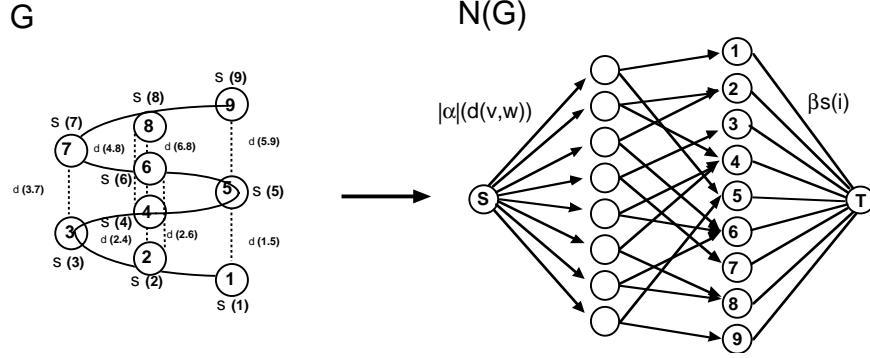


Figure 1: Contact Graph to Network Flow Graph Conversion

Definition (H-assignment): Let $G_C = (V, E, s, d)$ be a contact graph. An element $H \in \wp(V)$ is called an *H-assignment*.

Definition (energy): Let $G = (V, E, s, d)$ be a contact graph. $E_G: \wp(V) \rightarrow R$ is the *energy function* for G , where $E_G(H) = -\alpha \sum_{e \in E} d(e) + \beta \sum_{v \in H} s(v)$ and $E' = (H \times H) \cap E$.

Definition (optimal H-assignment): An optimal assignment for a contact graph G is an H-assignment X , such that $E(X) = \min\{E_G(X') \in R \mid X' \text{ is an H-assignment}\}$.

Definition (GC-IPF):

Given: A contact graph G_C

Find: An optimal H-assignment for G_C

2.2 Algorithms for GC-IPF

Herein, we review the conceptual basis for Kleinberg's approach and introduce an improved algorithm. Let $G_C = (V, E, s, d)$ be a contact graph.

Overview of Kleinberg's algorithm:

1. The contact graph G is transformed into a bipartite network flow graph $N(G) = (V', E')$, where $V' = \{s, t\} \cup E \cup V$, $s, t \notin E \cup V$ and $E' = (\{s\} \times E) \cup (V \times \{t\}) \cup \{(e, v) \in E \times V \mid \exists w \in e: \{v, w\} \in E\}$. The capacity for an edge $(s, e) \in E'$ is $\alpha d(e)$, the capacity for an edge $(v, t) \in E'$ is $\beta s(v)$, and all remaining capacities are infinite. An example for $N(G)$ is shown in Figure 1.
2. Find a minimum cut $C = (X, Y)$ in $N(G)$, where $s \in X$, $t \in Y$.

It can be shown that the H-assignment $H = X \cap V$ is an optimal assignment [18].

To describe the running time we define $n=|V|$ and $m=|E|$. Step 1 needs $O(n+m)$ time and the running time of step 2 depends on the network flow algorithm used. Kleinberg applies an algorithm designed for general flow networks [21,22]. This approach solves step 2 in $O((n+m)^2 \log(n+m))$ time, which is also the overall running time. Modification of Kleinberg's approach: The network flow algorithm used is designed for general flow networks, but the network flow graph that is calculated in step 1 is bipartite. For a bipartite network flow graph $B = (X, Y, E_B)$ a minimal cut can be found in $O(n_x m \log(n_x/m_B))$, where $n_x=|X|$, $n_y=|Y|$, $x \leq y$, and $m_B=|E_B|$ as shown by Ahuja et al. [23]. Thus by using the Ahuja et al. algorithm in step 2 GC-IPF can be solved in $O(nm \log(n/m + 2))$ time.

Kleinberg [18] assumes that in practice volume constraints in three dimensions imply that the number of possible contacts has a constant upper bound. From this assumption follows the running time of $O(n \log n)$ for Ahuja et al., algorithm and $O(n^2 \log n)$ for Kleinberg's algorithm.

2.3 Optimal Assignments: Structure and Algorithm

Clearly, there can be 2^n different optimal H-assignments for a contact graph $G_C = (V, E)$, where $n=|V|$. In practice, it is most unlikely that this would result when the input is based on a contact graph for a PDB structure. As shown by Picard et al. [19], minimal cuts in a network flow graph are structured.

Theorem: If (S, \bar{S}) and (S', \bar{S}') are minimal cuts in a flow network, then $(S \cap S', \overline{S \cup S'})$ and $(S \cup S', \overline{S \cap S'})$ are also minimal cuts.

As stated in section 2.2, minimal cuts represent optimal assignments. Thus, for all optimal assignments the assignments with minimal and maximal cardinality, denoted as H_{min} and H_{max} , are unique. All optimal assignments are contained in H_{max} and contain H_{min} .

Biologically this implies that there exists a core pattern of H-assignments that must be present in all of the sequences mapping to a particular structure, and that evolutionary forces will be constrained to maintain this pattern so long as the structure does not change. Conversely it implies that changes in structure can have significant effects on evolutionary opportunity. This is because the constraints associated with maintaining a particular protein structure affect the shape of clouds in protein space thereby affecting MED. Picard et al. [18], proved that all minimal cuts can be computed linear in the output size, when the maximal flow is given.

3. Implementation

In this section we describe details of our implementation for the MED problem. Our implementation takes as input a file from the Protein Data Bank (PDB) [24]

containing the 3D coordinates of each C_α atom in a protein, and outputs the set of optimal HP sequences.

3.1 Parameters and Performance

Parameters: For a contact graph $G_c=(V,E,s,d)$ we specify the distance function d , the solvent accessibility s , and the scaling parameters α and β , following Sun et al. [1]. The distance function d is given by the following equation:

$$d(\{i,j\}) = \frac{1}{1 + e^{(6.5\text{\AA} - d_{ij})}} \quad (3.1)$$

The value d_{ij} is the distance between the C_α atoms of the amino acids corresponding to the nodes i and j in the contact graph G . The function s is given by DSSP [25]. The range of α is provided by the user, while $\beta=1/3$.

Performance: The algorithms were implemented in C++ and compiled with Visual C++ 6.0 using the standard debug mode. We ran our software on a 700 MHz PC with 256 MB of RAM under Windows 2000 Professional Edition. The running time for each of the 15 globin structures was usually less than three minutes.

4. Application to Real Data

In this section we report results for a trial of our approach. We applied the method to a *test set* of 15 globin structures taken from the PDB (Table 1). The test set was chosen to represent a diverse sampling of structures within a protein family and included myoglobins, leghemoglobins, clam hemoglobin, ferric hemoglobin and several hemoglobins from the following animals: human, goose, turtle, trout, shark, skate, lamprey, and the sea cucumber. Selection was restricted to monomeric globins and alpha chains of multimeric globins.

4.1 A Test Run Using Globins

We obtained a total of 19 optimal HP-sequences when we subjected the test set to the GC model at near optimal α values. The number of sequences generated from each structure varied from 1 for several of the taxa to 4 for soybean leghemoglobin. The optimal α values were estimated by comparing α values from 1 to 15 at 0.5 increments. We converted the amino acid sequences associated with each structure in the test set to their HP equivalents following Sun et al. [1]. The resulting HP strings are denoted *transformed sequences*. To explore the effect that changing α had on the output of the model we explored two parameters: The accuracy and the

energy difference. The accuracy is measured as the percent sequence similarity between the estimated and transformed sequences and the energy difference is measured as the difference between the energy of the estimated and transformed sequences. When we plotted both parameters against α we found that the energy difference behaved in a concave manner, while the accuracy behaved convexly.

We compared the set of sequences obtained from the GC model with the HP transformed sequences taken from nature for each of the 15 conformations. Our results indicate that many of the estimated assignments are close to those of the transformed sequences. In general the real sequences exhibit a finer grained sequence variation than the HP sequences generated by the GC model. Sequences produced by the GC model tended to have larger uninterrupted "runs" of H's and P's than were observed in the real sequences (sequences available on request).

The accuracies we obtained at optimal α values for each of the 15 globin structures were comparable to those obtained by Sun et al. ranging from 66.4 to 78.4% (Table 1).

Table 1: Globin structures used in test case. The number of optimal HP sequences is indicated in column 4. Percent accuracy reflects the similarity between the optimal HP-sequences and their counterpart transformed sequences.

PDB ID	Structure	Species	# seqs.	% accy
1A3N	human hemoglobin	<i>Homo sapiens</i>	1	66.7
1HV4	goose hemoglobin	<i>Anser indicus</i>	1	67.4
1OUT	trout hemoglobin	<i>Onchorhynchus mykiss</i>	1	69.0
1GCV	shark hemoglobin	<i>Mustelus griseus</i>	1	71.1
1CG5	stingray hemoglobin	<i>Dasyatis akajei</i>	1	70.2
3LHB	lamprey hemoglobin	<i>Petromyzon marinus</i>	1	70.5
1LHT	sea turtle myoglobin	<i>Caretta caretta</i>	1	77.8
1VXB	sperm whale myoglobin	<i>Physeter catodon</i>	1	78.4
1GDJ	yellow lupine leghemoglobin	<i>Lupinus luteus</i>	2	72.2
1BIN	soybean leghemoglobin	<i>Glycine max</i>	4	66.4
1MOH	clam ferric hemoglobin	<i>Lucina pectinata</i>	1	72.5
1BOB	clam hemoglobin	<i>Lucina pectinata</i>	1	73.2
1HLM	sea cucumber hemoglobin	<i>Caudina arenicola</i>	1	70.9
1 H97	Trematode hemoglobin	<i>Paramphistomum</i>	1	77.6
1DLY	unicellular alga hemoglobin	<i>Chlamydomonas</i>	1	70.2

The ClustalW [26] alignment of the 19 optimal HP-sequences was subjected to UPGMA cluster analysis. Optimal HP-sequences clustered according to the structure from which they were derived as shown in Figure 2. These results imply that although the various different globins are structurally similar, the minor

structural differences that do exist are sufficient to affect the set of optimal HP-sequences.

This observation is consistent with the theoretical prediction described in section 2.3. This finding has potentially important implications for phylogenetic analysis. If minor differences in protein structure are responsible for significant changes in constraints in nature, then the models of evolution used to estimate phylogenetic trees from naturally occurring sequences must incorporate the changes if they are to avoid yielding misleading results. Put another way, if the models used to estimate phylogenetic trees do not accommodate the non-stationarity in the process caused by constraint changes, the resulting trees will be misleading.

The relationships among the 15 clusters of HP sequences resulting from the application of CG-IPF, mirrored the cluster structure obtained when each of the transformed sequences was examined using a cluster analysis (Figure 2). This adds further weight to the idea that the representation we used is capturing something natural about the mapping between sequence and structure. While we acknowledge that considerable differences exist between the optimal HP-sequences and their transformed counterparts, we find it interesting that the patterns of similarity among the HP-sequences show such correspondence to that of the real sequences.

4.2 Exploring Sub-Optimal α Values

We computed optimal HP-assignments at 90% and 80% of the optimal energy using a α value of 0.1. As expected, the number of optimal HP-sequences increased as the optimal energy criterion was relaxed. 40 optimal HP-sequences resulted at 90% of the energy while 48 resulted at 80%. UPGMA cluster analysis of these sequences yielded a similar cluster structure to that obtained with the optimal energy (Figure 2). Note that the sequences generated at near optimal energy values are a subset of those generated at sub-optimal energy levels. We did not encounter any cases where sequences from different structures clustered together to the exclusion of sequences from the same structure. Presumably, there would be a point at which clusters start to overlap. This point would represent the "bridge point" at which different protein structures become inter-accessible.

4.3 Scope of the Model

This approach and our implementation of it face several potential problems and limitations. Our results show that our model tends to over specify the problem to the point that only a few optimal HP sequences are produced out of an exponential potential number of possible mappings. This problem, however, is tied directly to our implementation and can be solved by reducing the specificity of the distance and solvent accessibility functions.

Our approach is also limited by a dependency on a distance measure for comparing sequences between proteins. Finding an appropriate distance measure for comparing diverse proteins is a challenging problem that has not yet been adequately solved.

5. Outlook

Sequences that map to a particular protein structure form a cloud of points in protein space. The shape of such clouds reflects the structural and functional constraints of the protein [18]. As structures and functions change over the course of evolution, so does the shape of the corresponding cloud. When one cloud comes close to another, the two protein structures represented by the clouds, become evolutionarily inter-accessible. If a mutation occurs that allows a sequence in one cloud to be converted into a sequence in another, a conformational change in the protein will result. Recent empirical evidence suggests that these conversions occur in nature [27,28,29,30].

The protein space representation provides a clear insight into the redundancy of the gp-map. It explains how resilience in the phenotype can be reconciled with a genotype that is free to explore different configurations [31]. Redundant mappings promote evolutionary access to new phenotypes and are predisposed to yield discrete changes at the phenotype level. We conjecture that some of the abrupt morphological changes seen in the fossil record represent discrete transitions caused by redundant mappings at higher levels of biological organization. An understanding of the gp-map at the organismal level will likely remain beyond our reach for some time to come. However, we believe that advances in algorithmic approaches combined with representations that meaningfully capture biology at the molecular level, promise to bring an understanding of the gp-map at the protein level within grasp in the near future.

Acknowledgements

We are grateful to Cecilia Clementi for pointing out empirical work consistent with the predictions of quantum evolution in protein space. During the review process we learned that a similar approach had been explored independently by Aspnes et al. [32] at Yale University. Their contribution is to appear in ISAAC. NSF grant DEB 9707145 to GJPN and DEB 0075319 to OE supported the current work.

References

1. S. Sun, R. Brem, H. Chan and K. Dill, *Protein Eng.*, **8**, 12, 1205-1213 (1995)
2. E.I. Shakhnovich and A.M. Gutin, *Protein Eng.*, **6**, 8, 793-800 (1993)
3. L. Holm, C. Sander, *Science*, **273**, 595 (1996)
4. P. Schuster, W. Fontana, P.F. Stadler and I.L. Hofacker, *Proc. R. Soc.*, **255**, 279-284 (1994)
5. H. Li, R. Helling, C. Tang and N. Wingreen, *Science*, **273**, 666-669 (1996)
6. G.J.P. Naylor and M. Gerstein, *J. Mol. Evol.*, **51**, 223-233 (2000)
7. D.J. Lipman and W.J. Wilbur, *Proc. R. Soc.*, **245**, 7-11 (1991)
8. J. Maynard Smith, *Nature*, **225**, 563-564 (1970)
9. C. Reidys, P.F. Stadler, and P. Schuster, *Bull. Math. Bio.*, **59**, 339 (1997)
10. M. A. Huynen, P.F. Stadler and W. Fontana, *PNAS*, **93**, 397-401 (1996)
11. K.A. Dill, S. Bromberg, K. Uayue, K. Fiebig, D. Yee, P. Thomas and H. Chan, *Protein Science*, **4**, 561 (1995)
12. K. Yue and K.A. Dill, *PNAS*, **89**, 4163-4167 (1992)
13. J.M. Deutsch and T. Kurosky, *Phy. Rev. Lett.*, **76**, 2, 323-326 (1996)
14. W. Hart, *Proc. RECOMB*, 128-136 (1997)
15. S. Kamtekar, J. M. Schiffer, H. Xiong, J.M. Babik and M.H. Hecht, *Science*, **262**, 1680-1685 (1993)
16. E.I. Shakhnovich and A.M. Gutin, *PNAS*, **90**, 7195-7199 (1993)
17. Lau, K.F. and K.A. Dill, *Macromolecules*, **22**, 3986-3997 (1989)
18. J.M. Kleinberg, *Proc. RECOMB*, 226-2367 (1999)
19. J. Picard and M. Queyranne, *Math. Prog. Study*, **13**, 8-16 (1980)
20. R. Ahuja, T. Magnanti and J. Orlin in *Network Flows*, (Prentice Hall 1993)
21. D. Sleator and R.E. Tarjan, *JCSS*, **26**, 362-391 (1983)
22. A.V. Goldberg and R.E. Tarjan, *Journal of ACM*, **35**, 921-940 (1988)
23. R.K. Ahuja, J.B. Orlin, C. Stein and R.E. Tarjan, *SIAM*, (1994)
24. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, *Nuc. Acids Research*, **28**, 235-242 (2000)
25. W. Kabsch and C. Sander, *Biopolymers*, **22**, 2577-2637, (1983)
26. J. D. Thompson, D. G. Higgins and T. J. Gibson, *Nuc. Acids Research*, **22**, 4673-4680 (1994)
27. R.D. George, *Nature Struct. Biol.*, **4**, 7, 512-514, (1997)
28. S. Dalal, S. Balasubramanian and L. Regan, *Nature Struct. Biol.*, **4**, 7, 548-552 (1997)
29. S. Dalal and L. Regan, *Protein Science*, **9**, 9, 1651-1659 (2000)
30. M.H. Cordes, R.E. Burton, N.P. Walsh, C.J. McKnight and R.T. Sauer. *Nature Struct. Biol.*, **7**, 12, 1129-1132 (2000)
31. B. Küppers, *Molecular Theory of Evolution*, (Springer-Verlag, NY 1983)
32. J. Aspnes, J. Hartling, M.Y. Kao, J. Kim, and G. Shah in *Lecture Notes in Computer Science: 12 Ann. Inter. Symp. on Algorithms and Computation*, Eds. P. Eades and T. Takaoka, (Springer-Verlag, NY 1983)

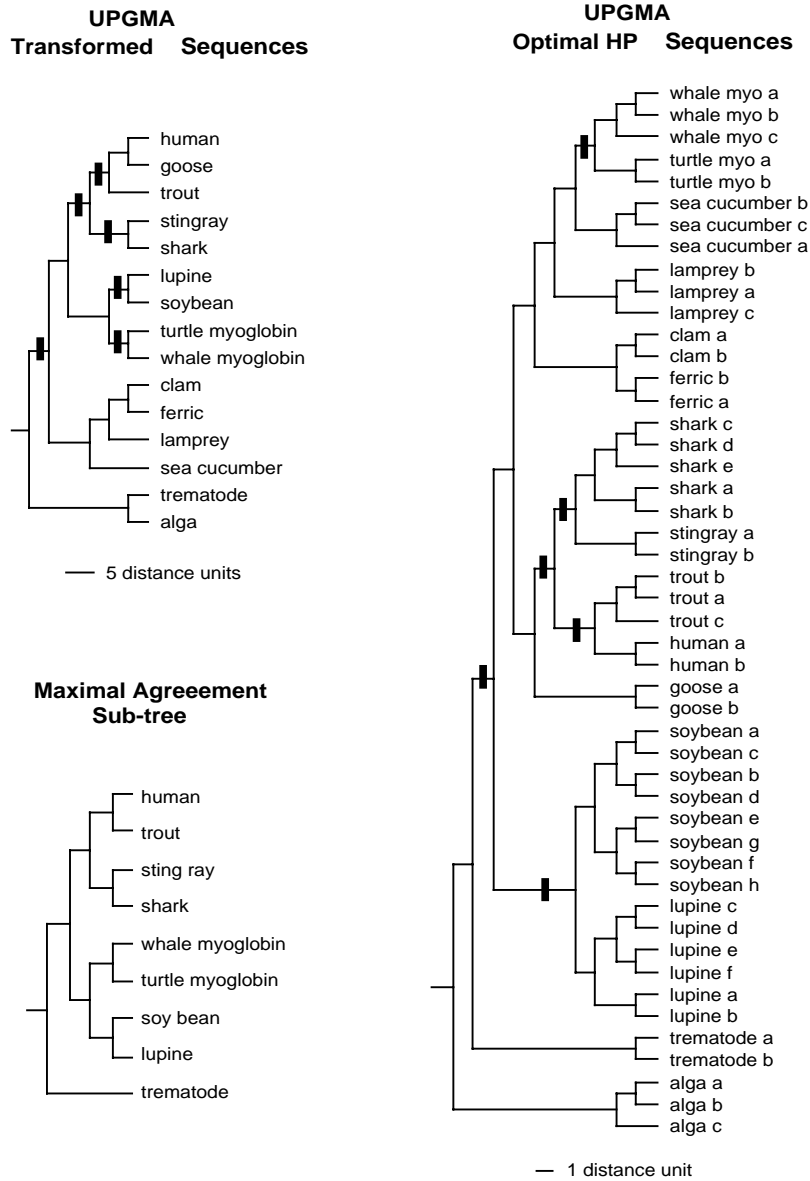


Figure 2: UPGMA of 48 optimal HP assignments resulting from the application of the GC-IPF to 15 globin structures (right). The UPGMA for the transformed sequences is shown at top left. The Maximal Agreement sub-tree shows groupings with a common cluster structure between the two analyses.