

# Accurate Detection of Very Sparse Sequence Motifs

ANDREAS HEGER,<sup>1</sup> MICHAEL LAPPE,<sup>2</sup> and LIISA HOLM<sup>3</sup>

## ABSTRACT

**Protein sequence alignments are more reliable the shorter the evolutionary distance. Here, we align distantly related proteins using many closely spaced intermediate sequences as stepping stones. Such transitive alignments can be generated between any two proteins in a connected set, whether they are direct or indirect sequence neighbors in the underlying library of pairwise alignments. We have implemented a greedy algorithm, MaxFlow, using a novel consistency score to estimate the relative likelihood of alternative paths of transitive alignment. In contrast to traditional profile models of amino acid preferences, MaxFlow models the probability that two positions are structurally equivalent and retains high information content across large distances in sequence space. Thus, MaxFlow is able to identify sparse and narrow active-site sequence signatures which are embedded in high-entropy sequence segments in the structure based multiple alignment of large diverse enzyme superfamilies. In a challenging benchmark based on the urease superfamily, MaxFlow yields better reliability and double coverage compared to available sequence alignment software. This promises to increase information returns from functional and structural genomics, where reliable sequence alignment is a bottleneck to transferring the functional or structural characterization of model proteins to entire protein superfamilies.**

**Key words:** protein sequence alignment, consistency, protein sequence motifs, transitive alignment.

## 1. INTRODUCTION

**P**ROTEINS CAN BE CLUSTERED INTO FAMILIES based on statistically significant sequence similarity. Protein families can be further clustered into structurally conserved and functionally related superfamilies. Superfamilies unify remote homologues whose similarity is difficult to detect using current sequence comparison methods, and the unification is often based on structure comparison. In addition to the difficulty of detecting homology between remote homologues, reproducing the structural alignment of remote homologues using only sequence information is a difficult challenge to existing multiple alignment or motif search software and is essentially an unsolved problem.

Multiple sequence alignment is an increasingly important problem in the era of genomics, because reliable alignments are required as input to programs for identifying functional sites and model building by

---

<sup>1</sup>Institute of Biotechnology, University of Helsinki, Finland.

<sup>2</sup>EMBL-EBI, Cambridge, United Kingdom.

<sup>3</sup>Institute of Biotechnology and Department of Genetics, University of Helsinki, Finland.

homology. Common wisdom holds that reliable alignments are obtained only if sequence identity is above 20–40% (Sander and Schneider, 1991; Lindahl and Elofsson, 2000). However, structure comparisons have revealed many large and diverse superfamilies where sequence identities between distant members are as low as 10% (Dietmann *et al.*, 2001). For example, alignment accuracy is one of the focus areas of the CASP competition (critical assessment of structure prediction).

We were particularly motivated to the present work by the failure of available sequence alignment programs to deliver correct alignments of the complete active site in the urease superfamily, which we have previously studied in detail (Holm and Sander, 1997). The structural alignment identifies invariant residues of the active site (Fig. 1). Such signatures are diagnostic of more distant homologues which

```

MKINRQQAESYGPTVGEVRLADTDLWIEVEKDYTTYGDEVNFGGKVLREGMGNGTYTRTENVLDDLTLNAILDYT 1UBPC
-----ALQTIINARLP----- 1K6WA

-----SQVLKIRRPDDWHLHLRDG----- 1J79A
GIYK-ADIGVKDGYIVIGIGKGNPDIMDGVTPNMIVGTATEVIAAEGKIVTAGGIDTHVHFI----- 1UBPC
-----DRINTVRGPITISEA-GFTLTHEHICGS----- 1PSCA
GEEGLWQIHLQDGKISAIQAQSGV-----MPI-TEN-SLDAEQGLVIPPVPEPHIHLDTTQTAGQP----- 1K6WA
-----TPAFNKPVELVHVLHGAIKPETILYFGKRGIA 1A4MA
* *
-----D-----MLKTVPVYTS- IYGRAIVM----- 1J79A
-----NPDQVDVALANGITTLFGGGTGPAEGSKA 1UBPC
-----SAGFLRAWPEFFGSRKALAEKAVRGLRRARAAGVRTIVDV----- 1PSCA
-----NWNQ--S--GTLFEGIERWAERKALL-----THDDVKQRAWQTLKWQIANGIQHVRTH----- 1K6WA
LPADTVEELRNIIGMDKPLSLPGFLAKFDYMPVIA-----GCREAIKRIAYEFVEMKAKEGVVYVEVR----- 1A4MA

--PNLAPPV-----TTVEAAVAYRQRI-LDAVPAPHDFTPMLTCYLTD-----SL-----DPNELERG 1J79A
TTVTP-----GPWNIKMLKST-EGLP-----INVGILGKGGH-----S-----SIAPIMEQ 1UBPC
--STFD-IG-----R--DVSLLAEVSRAAD-----VHIVAATGLWFDPPPLSMRLRSVEELTQFFLRE 1PSCA
--VDVSD-----ATLTALKAMLEVK-QEVAP--WIDLQIVAFPQEGIL--SY-----PNGEALLEEA 1K6WA
--YSPHLLANSKVDMPWNQTEGDVTPDDVVDLVNQGL-QEGEQA-FGIKVR SILCCMRHQ-----SWSLEVLEL 1A4MA

F-NE-----GV--FTAAXLYP-ANATTNSSHGVTSD-AIMPVLERMEKIGMPLLHGE-VTHADIDIFD-REARFIES 1J79A
I-DA-----G--AAGLXIHE-DW-----GATPA-SIDRSLTVADEADVQVAIHSD-TLNE-----AGFLE- 1UBPC
I-QYGIEDTGIR--AGIKVAT-TGK--A--TPFQEL-VLKAARASLATGVPVTTHTA-A-----SQRDGE- 1PSCA
L-RL-----G--ADVVGAI P-HFEFT--REYGV E-SLHKT FALAQYDRLLIDVHCDEI-----DDEQ--SRFVE- 1K6WA
CKKY-----NQKTVVAMDLAGDETI-E-----GSSLFPGHVEAYEGAVKNGIHRTVHAGEV-----GS--PEVVR- 1A4MA
*
VMEPLRQRLTAL--KVFEHIT-----K-----DAADYVRDGNERLAATI TPQHLMF--NRNHMLVGGV RPHLY 1J79A
DTLRAIN G--R--VIHSFHVEGAGGGHAP-----DIMA-MAGHP-NVLPSTNPTRFPTVNTIDEH----LDMLM 1UBPC
QQAAIFESEGLSPSRVCIGHSD-----TDD-----LSYLTALAA--RGLYIGLDH-IPH--SAIGLED--NASASA 1PSCA
TVAALAHHEG-MGARVTASHTTA-----MHSYNGAYTSRLFRLLKM--SGINFVANPLVNIH----- 1K6WA
EAVDILK-----TERVGHGYH-----TIE-----DEALYNRLK--ENMHFEVCPWSSYL----- 1A4MA
*
CLPIL-----KR-----NIH-QQALRELVASG-FQRVFLGTDSAPHA--RHRKESSCGCAGCF-NAPTAL 1J79A
-VCHHLKQNIPEVAFADSRIR----PET-IAAEDILHDLG-I-ISMMSTDALA-----MG-RAGEMV 1UBPC
L--LG-----IR--SQTR-ALLIKALIDQGYMKQILVSNLWLPGFSSYVTNIMDVMDRVNPDGMAFIP 1PSCA
LQ-----GRFDTPKRRGITR.VKEMLESG--INVCFGHDDVFD-----P--WYPLGTA-NML 1K6WA
TGA-W-----DP-----KTT--HAVVRFKNDK--ANYSLNTDDPL-----IF-KS--TLD 1A4MA
*
GSYATVFEEMN-----ALQ-HFEAFCSVNGPQFYGL-----PVND-----TFIELVR-EEQQVAESI 1J79A
LRTWQTADKMKKQQRGPLAEKNGSDNFR LKRYVSKYTINPAIAQGI-----AHEVGSIEEGKFADLVLWEPKF----- 1UBPC
LRVIPFLREK-----VPQETLAGITVTNPARFLS----- 1PSCA
QVLHMGLHVCQL--MGY-----GQINDGLNLITTHSARTL-N--LQDY-GIAAGNSANLILPAE----- 1K6WA
TDYQMTKKDMG-----FTEEFKRL-NINAAKSFLPEEKKELLERLYRE---YQ----- 1A4MA

ALTDDTLVPFLAGETVRSVK----- 1J79A
-----FGVKADRVIKGGIIAYAQIGDPSASIP TPQVPMGRMYGTVDGLIHDTNITFMSKSSIQQGVPAK 1UBPC
----- 1PSCA
-----NGFDALRRQVPVRYSVRGGKVIAS T-----QPAQ--TTVYLEQPEAIDYKR----- 1K6WA

LGLKRRIGTVKNCRNIGKDMKWNVDVTTDIDINPETYEVKVDGEVLTCEPVKELPMAQRYFLF 1UBPC
-----PTL-R----- 1PSCA

```

**FIG. 1.** Structural alignment (Dietmann *et al.*, 2001) of five members of the urease superfamily. The challenge is to detect the active site signature motif (asterisks) using only sequence information. The signature positions are located at the ends of beta strand 1 (*H.H*), 4 (*H*), 6 (*H*), and 8 (*D*) of a (beta/alpha)<sub>8</sub>-barrel.

can be pulled out from databases based on sequence similarities. To date, we have identified 26 main branches (families) in the phylogenetic tree of the urease superfamily. A number of recent structure determinations of proteins representing different parts of the urease superfamily have verified earlier blindfold predictions.

The signature motif of the urease superfamily can be recognized by eye in multiple alignments of each family, but all automatic programs—we have tried many over the years—have failed to deliver a complete alignment across the whole superfamily that would preserve the correct register of the signature motif between families. Here, we present MaxFlow, the first program to reproduce the essential parts of the structural alignment, recognizing correctly 18 out of 20 motif positions marked by asterisks in Fig. 1.

## 2. THEORETICAL BACKGROUND

Protein sequence evolution is usually modeled using simple statistical models of amino acid preferences at a given position. A sequence is aligned to the model (profile), and the statistical significance of the alignment score is estimated. Thus, one can both detect homologues and obtain a sequence-profile alignment with this approach. The profile model can be derived for one sequence (using amino acid substitution matrices as in Blast, Fasta, SSearch), or it can be learned from a set of known homologues (as in MaxHom [Sander and Schneider, 1991], PSI-Blast [Schaffer *et al.*, 2001], hidden Markov models [Eddy, 1998], Gibbs sampling [Lawrence *et al.*, 1993], MEME [Grundy *et al.*, 1997]). The result of a database search is an effective multiple alignment of many sequences against each other, via the profile.

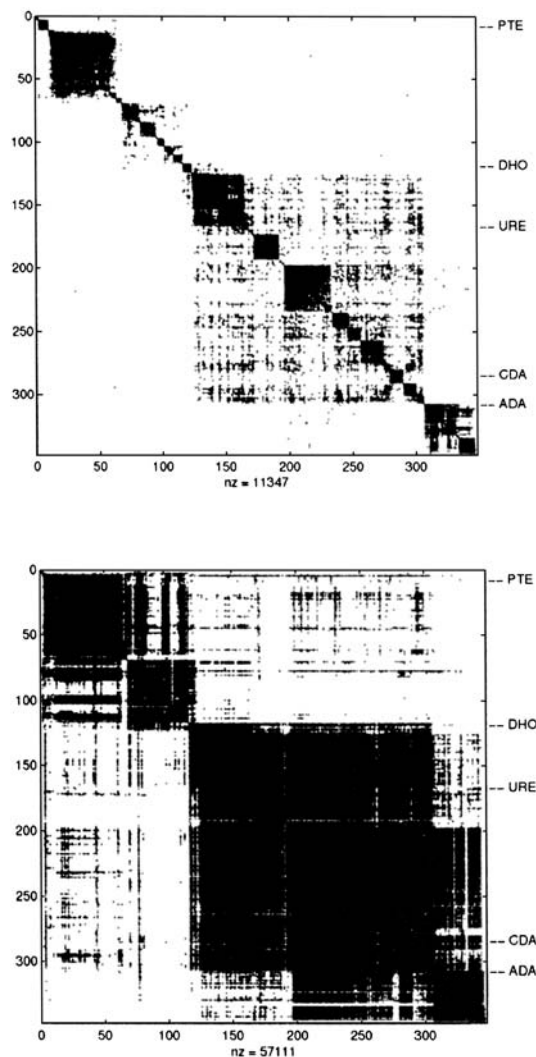
Programs like PSI-Blast are powerful and popular, but there is an intrinsic limitation to the profile approach. Profile models are based on log-odds scores for observing a given amino acid type at a given position. From an information theoretical perspective, profile scores “dilute out” as the target distributions of amino acid types observed in structurally equivalent positions broaden at large evolutionary distances and approach the background distribution (Altschul, 1991). For example, there is hardly any sequence conservation around the invariant histidines of the urease signature motif (Fig. 1). While sparse signature patterns like the ones in Table 1 are diagnostic of enzyme active sites in these particular superfamilies, they have too low information content to be discriminative in database searches (i.e., almost any histidine-rich protein will match the urease signature pattern).

From an evolutionary perspective, the profile model is centred at one point in protein space and will detect a specific subset of sequence neighbors at some limited radius. However, proteins in different parts of a large superfamily evolve under a different set of functional constraints, which affect amino acid preferences and which set of proteins will be detected with statistically significant similarity. We used PSI-Blast to generate a global “map” of the urease superfamily (Fig. 2). PSI-Blast profiles were seeded at each protein, and dots indicate significant similarity to the profile model (sequence neighbors). There is no PSI-Blast profile that would recognize all members of the superfamily. Expert classifications of protein families like Pfam also use several models to describe different parts of the superfamily. However, all proteins are connected directly or indirectly to each other by PSI-Blast alignments. Thus, the pairwise alignment library underlying Fig. 2 implies a transitive alignment between, for example, the first and last protein via a number of intermediates.

TABLE 1. TRACE GRAPHS<sup>a</sup>

<i>Superfamily</i>	<i>Source</i>	<i>N</i>	<i> V </i>	<i> E </i>	<i>Active site signature</i>
Urease	Blast PSI-Blast	347	150,673	1,044,049 12,085,888	<i>H.H.</i> {110, 197} <i>H.</i> {24, 38} <i>H.</i> {57, 88} <i>D</i>
DNA polymerase $\beta$	PSI-Blast	107	37,074	1,848,314	<i>G.</i> {10, 11} <i>D.</i> [ <i>DE</i> ].{23, 85}[ <i>DE</i> ]
Actin	PSI-Blast	206	90,226	4,808,517	<i>D.G.</i> {96, 257} <i>G.</i> {20, 67} <i>G.</i> {88, 113} <i>G.</i> {26, 40} <i>G</i>

<sup>a</sup>*N*, number of representative sequences in superfamily; *|V|*, number of residues; *|E|*, number of aligned residue pairs in alignment library. In the signature patterns, [ ] denote alternative amino acids, and { } denote the lower and upper limit on the length of the intervening sequence in known structures from the benchmark set (Table 2).

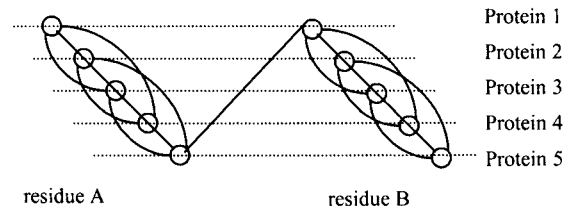


**FIG. 2.** Symmetrized neighbor matrix of 347 representative proteins from the urease superfamily. Dots mean PSI-Blast  $e$ -value  $< 1$ . **Top:** first iteration (ordinary Blast search); **bottom:** last iteration ( $10^{\text{th}}$  unless converged).

The problem with transitive alignments of course is that there are very many choices of intermediates that will lead to mutually inconsistent alignments between the proteins at the start and end of the chain. In the classical multiple sequence alignment problem, one aims to reconcile such inconsistencies using ad hoc objective functions, usually a sum-of-pairs score (Kececioglu, 1993). This problem is NP-complete, and therefore practical algorithms for multiple alignment of large sequence sets have to cut corners or make drastic assumptions (Notredame, 2002). In this work, we propose a novel type of objective function for transitive alignment which is based on a path score. The path score is defined between any two proteins in a connected set, even if there is no direct pairwise alignment between the proteins in the alignment library. Thus, we can compare two very remote homologues but need address only the pairwise alignment problem, which can be solved exactly.

### 3. OBJECTIVE FUNCTION

We use a library of pairwise alignments to generate an alignment trace graph. The trace graph  $G = (V, E)$  is an undirected unweighted graph derived from pairwise alignments of proteins by some method (such



**FIG. 3.** Schematic alignment trace graph. The vertices  $V$  are residues. Each residue is additionally labeled by the protein that it belongs to. The edges  $E$  represent all residue–residue pairings that occur in the pairwise alignment library.

as PSI-Blast) at a “safe” threshold (such as  $e$ -value  $< 1e-5$  in PSI-Blast, so that the alignments carry some evolutionary information). The nodes  $V$  of the alignment trace graph are all the residues in the input set of proteins. The nodes are labeled with the source protein and residue number. Edges  $E$  join each pair of nodes that have been aligned in the input library of pairwise alignments. Figure 2 represents a contracted version of a real trace graph (all residues from the same protein have been contracted). There is a statistically significant direct pairwise alignment between less than half of all protein–protein pairs in the urease superfamily but, importantly, the whole set is connected so one can find a path from any one protein to any other protein in this set. At the residue level, one can find a path from almost any residue to most other residues. As an objective function for transitive alignment, we define a consistency score to select the pairing of residues that has the most support in the alignment library (trace graph).

The objective function measures the consistency of alternative paths leading from one residue in the source protein to a residue in the target protein. For example, one can see from Fig. 3 that residue  $A$  in protein 1 is not directly aligned with residue  $A$  in protein 5, but these are connected through many alternative paths involving proteins 2–4. The direct pairwise alignment of residue  $A$  in protein 5 to residue  $B$  in protein 1 is actually not consistent with any of the other pairwise alignments in Fig. 3. Intuitively, one sees that the most consistent multiple alignment of proteins 1–5 is one that places all of residues  $A$  in one column and all of residues  $B$  in another column.

We define a local consistency score  $g(s, t)$  between adjacent nodes  $s$  and  $t$  in terms of common neighbors:

$$g(s, t) = \frac{|S(s) \cap S(t)|}{|S(s) \cup S(t)|}, \quad (1)$$

where  $S(x)$  is the set of direct neighbors of node  $x$  in the trace graph and  $|S(x)|$  is the cardinality of the set. This normalized consistency score has a range from zero to one. Using the consistency score as edge weights, we generalize this to define a path score for transitive alignment, where  $s$  and  $t$  need not be adjacent nodes in the trace graph. The formal definition follows

**Definition.** Let  $s$  be a residue in protein  $X$  and let  $t$  be a residue in protein  $Y$ , where  $X \neq Y$ . A valid path from  $s$  to  $t$  can visit the same protein only once. The path score is the weakest link along this path. The weakest link along a path is the pair  $(v, w)$  of neighboring nodes with minimal consistency score  $g(v, w)$ . The alignment score  $M(s, t)$  for pairing residue  $s$  with residue  $t$  is the maximal path score over all valid paths from  $s$  to  $t$ .

In other words, we search a graph for a path with the highest minimum. This appears original in sequence alignment context.

The notion of scoring alignments for consistency—rather than amino acid or nucleic acid similarity—has been around for a long time. Vingron and Argos (1991) applied an elegant matrix multiplication procedure to filter multiple dot-plots for motifs that occur consistently in all proteins in a given input set. Notredame (1999) proposed a novel objective function based on local triplet counts similar to Equation (1). In other words, one assigns higher confidence to an edge  $(s, t)$  in the alignment trace graph if there are many intermediates  $i$  connected to both  $s$  and  $t$  (i.e.,  $(s, i)$  and  $(i, t)$ ). The T-Coffee program (Notredame *et al.*, 2000) uses sophisticated heuristics to modify the edge weights and generates a multiple alignment based on the progressive strategy. In progressive alignment, sequences or groups of sequences (whose internal alignment remains fixed) are aligned to each other in the order determined by a phylogenetic tree. The novel

concept in MaxFlow is the path score, which extends over any number of intermediates in the trace graph. We will not here address the multiple alignment problem<sup>1</sup> but focus on *transitive pairwise alignments* that optimize the path score.

## 4. ALGORITHM

### 4.1. Outline

We treat transitive alignment as a clustering problem. We assume that each column of the true (that is, structurally defined) multiple alignment is present in the alignment trace graph as a connected subset of residues. Conceptually, our goal is to delineate these subsets (clusters) based on the consistency information extracted from the input library. If we wanted to solve the multiple alignment problem, we would have to eliminate all edges between structurally nonequivalent residues from the alignment trace graph. Here, we address the simpler problem of aligning just two proteins S and T. Moving along the links in the alignment trace graph, any residue from S can reach many residues in T, and vice versa. We use the path score to evaluate the relative likelihood of the paths leading to alternative residue pairings, in light of the information in the input library.

The definition of the path score in terms of the weakest link allows an efficient search over many alternative paths. The search makes use of hierarchical clustering. This divides the residues into subsets, and the path score between any members from two different subsets is simply the depth at which the two subsets are merged, without any need to specify an explicit path. The clustering is greedy and seeded from the most densely connected regions in the alignment trace graph. At this stage, residues may be assigned to clusters in a way which is not compatible with sequential ordering of residues in the final pairwise alignment. The greediness of the clustering is relieved by fanning out the search to a reasonably limited set of alternative clusters. The sequentiality issue is solved by dynamic programming in the last step.

### 4.2. Edge weighting

We first generate an alignment trace graph  $G$  from the input library of pairwise alignments. The edges in the alignment trace graph  $G$  are weighed according to the consistency score  $g(s, t)$  of Equation (1). Figure 4A illustrates the consistency score in a case where nodes  $s$  and  $t$  have four common neighbors (including self-matches) and three unique neighbors, yielding a consistency score  $g(s, t) = 4/7$ .

### 4.3. Greedy clustering

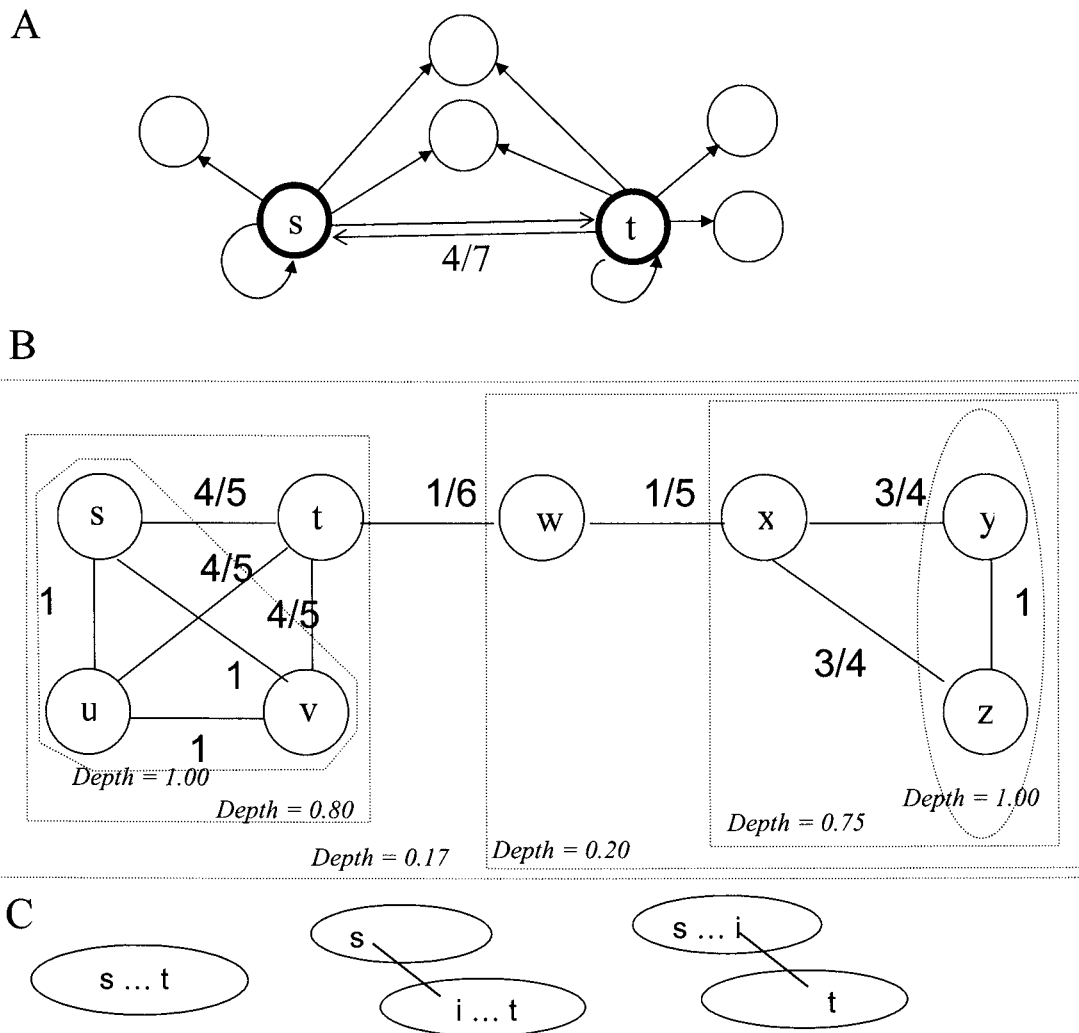
Nodes are assigned to clusters using Kruskal's algorithm (equivalent to single linkage clustering as illustrated in Fig. 4B), with the additional condition that each cluster may contain at most one residue from any one protein. The result is a disconnected graph  $H$ , where each connected component (cluster) contains one or zero residues from any one protein. If we consider only paths from residue  $s$  to residue  $t$  that visit residues within the same cluster, then the maximal path score  $h(s, t)$  is, by definition, equal to the depth at which  $s$  and  $t$  became merged in the clustering. Because of the greediness of the clustering, there may be a higher scoring path in the complete alignment trace graph. For this reason, the score matrix calculation (below) examines also paths that involve residues from two different clusters (Fig. 4C).

### 4.4. Score matrix

The path scores are computed for each residue pair  $s, t$  and stored in the scoring matrix  $M$  with dimensions  $m * n$ , where  $m$  and  $n$  are the lengths of the proteins.  $M(s, t)$  measures our relative confidence

---

<sup>1</sup>MaxFlow outputs pairwise alignments. These pairwise alignments are more consistent than the input alignment library but usually not perfectly consistent globally. If need be, one could quickly generate an explicit multiple alignment by combining pairwise MaxFlow alignments based on a tree with star topology (pile-up strategy) or a phylogenetic tree (progressive alignment). Iterated application of MaxFlow (using the output at iteration  $i$  as input to iteration  $i + 1$ ) should also lead to a self-consistent set of pairwise alignments that corresponds to a unique multiple alignment.



**FIG. 4.** **A.** The consistency score is defined in terms of neighbor overlap in the trace graph. For example, node *s* has four neighbors (including nodes *s* and *t*) in common with *t*. Node *s* has one unique neighbor, and node *t* has two unique neighbors. The consistency score is the ratio of the intersection and union of the neighbors of nodes *s* and *t*, that is, four out of seven. **B.** Schematic clustering of a simple alignment trace graph. Edge weights are defined as above. Dotted lines show the contour levels in hierarchical clustering by single linkage. The link strength between any two nodes within the same cluster is determined by the depth at which they are merged in the same subcluster. For example, the score between any node in the set {*s*, *t*, *u*, *v*} and any node in the set {*w*, *x*, *y*, *z*} is 0.17, and the score of node *x* with either node in the set {*y*, *z*} is 0.75. **C.** Each cluster contains at most one residue from any one protein. Here, clusters are represented as ellipses. To find the path with the highest weakest link, MaxFlow examines the case where nodes *s* and *t* are in the same cluster (*s*...*t*) or in different clusters but joined by an intermediate node in the original trace graph. Direct connections in the original trace graph are marked by dashes (*s*-*i* and *i*-*t*). The dotted connections *s*...*t*, *s*...*i*, and *i*...*t* denote paths which may involve many indirect neighbors in the original trace graph.

that residues *s* and *t* belong to the same column of the true (i.e., structure based) multiple alignment:

$$M(s, t) = \max_i \begin{cases} g(s, t) \\ h(s \dots t) \\ \min \left\{ \begin{array}{l} g(s, i) \\ h(i \dots t) \end{array} \right\} \quad \min \left\{ \begin{array}{l} g(t, i) \\ h(i \dots s) \end{array} \right\} \end{cases} \quad (2)$$

The term  $g(x, y)$  is defined iff residues *x* and *y* are directly linked in the alignment trace graph *G* and  $h(x \dots y)$  is defined iff *x* and *y* are assigned to the same cluster in *H*. The terms in the lower half of

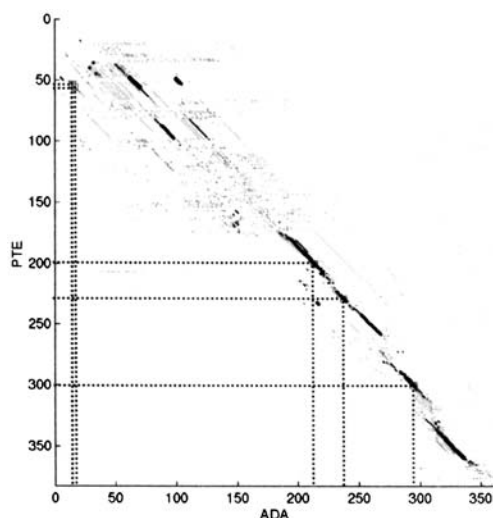
Equation (3) are evaluated only in the case that  $s$  and  $t$  are assigned to different clusters in  $H$ . (For example, if  $s$ ,  $i$ , and  $t$  all belong to the same cluster, then  $h(s, t)$  is by definition the minimum of  $g(s, i)$  and  $h(i \dots t)$  for any  $i$ .) In words, we check paths that connect  $s$  to  $t$  in the situation that  $s$  and  $t$  are directly aligned ( $s - t$ ,  $g(s, t) > 0$ ), in the same cluster ( $s \dots t$ ,  $h(s, t) > 0$ ), or they are in different clusters but one can make the connection via a path  $s \dots i - t$  or  $s - i \dots t$  which involves a residue  $i$  from an intermediate sequence and where “-” denotes a direct edge from graph  $G$  and “...” indicates a shortest path connection in the clustered graph  $H$ . Note that the dotted connection may join two residues which are indirect neighbors in the original alignment trace graph and separated by a large number of intermediates. We did not consider complex paths from  $s$  to  $t$  involving more than two clusters, as the search for valid paths would quickly become computationally prohibitive. Because of the precomputed multi-step paths ( $s \dots t$ ,  $i \dots t$ , or  $i \dots s$ ) in the clustered graph  $H$  from step 4.3, the search employed in this work is quite broad even with one intermediate  $i$ .

#### 4.5. Optimal alignment

The problem of aligning two proteins, given a scoring matrix, can be solved exactly using dynamic programming in  $O(mn)$ , where the sequences have lengths  $m$  and  $n$ . To derive the optimal alignment of proteins S and T, we fill the score matrix  $M$  for all residues  $s$  from S leading to all residues  $t$  from T. The matrix is sparse, as many entries are zero due to no connecting path (using the pruned search of step 4.4). We derived global alignments using the Needleman–Wunsch algorithm with zero gap penalties. This step is like a standard pairwise alignment, except that the score matrix is not derived from amino acid substitution tables but from MaxFlow path scores which have high information content despite very low amino acid similarities between the proteins being compared (Fig. 5).

#### 4.6. Implementation

The MaxFlow algorithm was implemented in C++. The program is available from the authors on request. The input to the program is a library of PSI-Blast alignments. We retrieved the alignment trace graphs precomputed from the PairsDb database (Heger, unpublished) which is implemented using Python and MySQL.



**FIG. 5.** Sequence derived score matrix  $M$  for the comparison of phosphotriesterase (PTE) and adenosine deaminase (ADA). PTE and ADA are indirect PSI-Blast neighbors from opposite extremes of the urease superfamily (cf. Fig. 2). Nevertheless, MaxFlow generates a strong alignment trace between these very distant homologues. Darker dots have higher scores. Dotted lines indicate the  $H.H-H-H-D$  motif. The motif positions lie on the optimal alignment trace, with the exception of the  $N$ -terminal  $H.H$  submotif which gets systematically misaligned in the PSI-Blast alignment library because it is followed by a long insert in ADA.

The edge weighting step (4.2) has complexity  $O(|E|N)$ , where  $N$  is the number of proteins in the input set. This step is straightforward to parallelize because of only local dependencies on the first neighbor shells of nodes  $s, t$  in the trace graph  $G$ . The residue clustering step (4.3), implemented using Kruskal's algorithm, has complexity  $O(|E| \log |E|)$ . The edge weighting and residue clustering steps are performed once for the whole superfamily. After this, any pairwise alignment can be generated by performing steps 4.4 and 4.5.

Step 4.4 fills the  $m * n$  score matrix for the pairwise comparison of protein sequences of length  $m$  and  $n$ .  $O(N)$  intermediates  $i$  are checked for each row and column, leading to complexity  $O((m + n)N)$ . The  $g(x, y)$  terms in Equation (3) were obtained by direct look-up. The topology of the single linkage trees for each component in graph  $H$  was stored in the residue labels as follows. The tree was cut into 100 slices at steps of 0.01 (the maximum possible consistency score is 1.0). The residue label records the branching of the tree at each slice; when two residue labels are compared, the length of their common prefix defines the depth at which they were merged. The  $h(x, y)$  terms were determined by the length of the common prefix in roughly constant time  $O(\log l)$ , where  $l$  is the length of the terminal node (i.e., residue) labels.

Step 4.5 uses standard dynamic programming which has complexity  $O(mn)$ . Our implementation is slightly faster in practice, because the score matrix is sparse and our alignment procedure visits only cells which have a positive score (dots in Fig. 5).

## 5. ALIGNMENT QUALITY ASSESSMENT

### 5.1. Small benchmark

We analyze in detail results from a small benchmark of structurally aligned proteins which are less than 25% sequence identical to each other and come from three superfamilies (Table 2). The structural alignment was generated by Dali (Dietmann *et al.*, 2001). Representative sequence sets of the superfamily were collected manually in the case of urease and from reciprocal PSI-Blast neighbors of the known structures in the other two superfamilies. Reciprocal PSI-Blast neighbors are sequence pairs  $X, Y$  where sequence  $Y$  is detected by a profile generated using sequence  $X$  as query (seed to the iterative profile

TABLE 2. STRUCTURALLY ALIGNED PROTEINS IN THE BENCHMARK SET<sup>a</sup>

<i>Abbr.</i>	<i>Description</i>	<i>Representative sequence</i>	<i>PDB code</i>	<i>Superfamily</i>
ADA	Adenosine deaminase	P00813	1a4mA	Urease
CDA	Cytosine deaminase	1246354	1k6wA	
URE	Urease	P07374	1ubpC	
DHO	Dihydroorotase	15155315	1j79A	
PTE	Phosphotriesterase	15212234	1pscA	
POL	DNA polymerase beta	1060896	1bpyA	DNA polymerase beta (catalytic domain)
TDT	Terminal deoxynucleotidyltransferase	Q99PD1	1jmsA	
PAP	Poly-A polymerase	P29468	1fa0B	
KAN	Kanamycin nucleotidyltransferase	O87369	1knyA	
GLU	Activator of hydroxyglutaryl-CoA dehydratase	Q9X5B6	1huxA	Actin (ATPase domain)
GLK	Glycerol kinase	Q9PB76	1glcG	
ACT	Actin-depolymerizing factor	Q9M351	1yagA	
FTS	ftsA	Q9WZU0	1e4fT	
HXK	Hexokinase	Q64476	1qhaA	
ACE	Acetate kinase	O83489	1g99A	
MRE	mreB rod shape determining protein	1572983	1jceA	
HSP	Hsp70	Q9W6Y1	1ba1	

<sup>a</sup>Representative sequences are identified by their Swissprot accession number (e.g., P00813) or Genpept identifier (e.g., 1246354).

construction) and sequence  $X$  is reciprocally detected by the profile generated using sequence  $Y$  as query. PSI-Blast was run with inclusion threshold 0.005 during iteration and hits were included if the  $e$ -value was  $< 1$  at convergence or up to the 10<sup>th</sup> iteration. The search database was NRDB40, a representative sequence database where all sequences are less than 40% identical to each other. MaxFlow was run using the representative sequences of the PDB structures. As the representatives are more than 40% identical to the structure, their sequence based alignment to the known structure is unambiguous.

### 5.2. Comprehensive benchmark

We also carry out comparisons in a comprehensive benchmark of structurally aligned proteins that cover the whole Protein Data Bank. All representative structures (at less than 25% sequence identity) in the Dali domain dictionary (Dietmann *et al.*, 2001) were organized in a spanning tree whose topology is defined by maximal  $Z$ -scores. The edges of the tree define the test pairs. Pairs that involved internal repeats were removed, leaving 669 structurally aligned pairs in the test set. Of these, 258 test pairs were direct PSI-Blast neighbors, and 420 pairs were PSI-Blast neighbors with 0–3 intermediates.

The input to MaxFlow is a set of sequences, which was selected from a “sequence space graph” based on the all-against-all PSI-Blast runs in NRDB40, a representative sequence database defined at a 40% sequence identity threshold. All Blast alignments with an  $e$ -value  $< 1e-5$  were included. The shortest path (up to a length of 4) between the test pair proteins was determined in the sequence space graph. Shortest paths up to a length of 4 were reported, else the test pair was treated as disconnected (about one third of the whole test set). The input set to MaxFlow included all the first, second, and third neighbors of the nodes along the shortest path.

### 5.3. Reliability and coverage

The quality of sequence-derived alignments was measured by reliability ( $TP/P$ ) and coverage ( $TP/T$ ). Here  $TP$  (true positive) is the number of correctly aligned pairs of residues,  $P$  (positive or predicted) is the total number of aligned pairs of residues in the sequence-derived alignment, and  $T$  (true) is the total number of structurally aligned pairs of residues in the Dali alignment.

We evaluate the accuracy of alignments at shift 0 or shift 0–5. The alignment shift is a measure of how close the sequence derived alignment is to the structural reference alignment. Shift 0 means that the structural alignment is reproduced exactly. It is common in the threading (protein structure prediction) community to classify aligned pairs as correct (acceptable) if the shift is  $\leq 5$ . A shift of four residues corresponds to one turn of an  $\alpha$ -helix and to second neighbors on the same side of a  $\beta$ -sheet. Let residue  $a_1$  from protein 1 be aligned with residue  $a_2$  from protein 2 in the sequence derived alignment. Let  $a_1$  be aligned with  $a_2^*$  and let  $a_2$  be aligned with  $a_1^*$  in the structural reference alignment. The alignment shift for  $(a_1, a_2)$  is defined as the minimum of  $|a_1 - a_1^*|$  and  $|a_2 - a_2^*|$ .

## 6. RESULTS

### 6.1. Urease superfamily

Reproducing the structural alignment of the urease superfamily using only sequence information is a difficult challenge to existing multiple alignment or motif search software (Table 3). MEME (Grundy *et al.*, 1997) searches for conserved blocks, but fails to identify any motif in DHO and mistakes the identity of the  $\beta - 5/\beta - 6$  histidine motifs of the urease superfamily, leading to less than perfect reliability. ClustalW (Thompson *et al.*, 1994) does progressive alignment optimizing a weighted sum of pairs. Comparison to pairwise profile–profile alignment (our own implementation, using profiles built from the Blast neighbors of the two query proteins) suggests that progressive is a very dangerous strategy for distantly related proteins. Dialign2 (Morgenstern, 1999) is a greedy segment-to-segment assembly algorithm, but also in this strategy errors made early on cannot be recovered from and the program loses the correct register between families. The T-Coffee program (Notredame *et al.*, 2000) was unable to handle the large sequence sets used here.

TABLE 3. ACCURACY OF SEQUENCE-DERIVED ALIGNMENT IN THE UREASE SUPERFAMILY BY DIFFERENT METHODS<sup>a</sup>

Method	Shift	TP												Total TP	Rel. TP/P	Cov. TP/T						
		ADA-URE		ADA-DHO		ADA-PTE		CDA-URE		CDA-DHO		CDA-PTE					URE-URE		URE-DHO		URE-PTE	
		ADA-URE	ADA-DHO	ADA-PTE	ADA-URE	ADA-DHO	ADA-PTE	CDA-URE	CDA-DHO	CDA-PTE	CDA-URE	CDA-DHO	CDA-PTE				URE-URE	URE-DHO	URE-PTE	URE-URE	URE-DHO	URE-PTE
3D alignment (T)	0	<b>276</b>	<b>219</b>	<b>227</b>	<b>238</b>	<b>294</b>	<b>239</b>	<b>258</b>	<b>247</b>	<b>219</b>	<b>238</b>	<b>2455</b>	<b>1.00</b>	<b>1.00</b>								
MEME	0	20	0	0	0	64	0	25	0	25	0	134	<b>0.55</b>	0.05								
ClustalW	0	14	3	7	8	8	9	46	20	8	13	136	0.06	0.06								
MaxFlow (Blast)	0	0	0	0	0	60	50	0	50	0	0	160	0.43	0.07								
Dialign2	0	69	17	18	38	39	35	15	39	17	47	334	0.33	0.14								
Profile-profile	0	99	0	0	0	120	85	9	73	8	9	403	0.49	0.16								
PSI-Blast	0	100	*56	0	0	119	<b>119</b>	0	46	0	0	533	0.32	0.22								
		<b>109</b>				145			104													
MaxFlow (PSI-Blast)	0	100	<b>67</b>	<b>65</b>	<b>95</b>	<b>151</b>	112	<b>127</b>	<b>106</b>	<b>79</b>	<b>98</b>	<b>1000</b>	0.39	<b>0.41</b>								
MEME	0-5	27	0	0	0	75	0	28	0	28	0	158	0.65	0.06								
ClustalW	0-5	32	17	20	38	19	39	125	102	17	32	441	0.19	0.18								
MaxFlow (Blast)	0-5	0	0	0	0	72	79	0	81	0	0	232	0.62	0.09								
Dialign2	0-5	106	47	36	57	70	60	40	53	37	72	578	0.57	0.24								
Profile-profile	0-5	134	0	0	0	162	138	9	118	8	9	478	0.58	0.19								
PSI-Blast	0-5	185	*122	0	0	216	<b>232</b>	0	158	0	0	1019	0.61	0.42								
		<b>200</b>				<b>261</b>			<b>204</b>													
MaxFlow (PSI-Blast)	0-5	193	<b>132</b>	<b>160</b>	<b>179</b>	211	203	<b>226</b>	158	<b>121</b>	<b>229</b>	<b>1812</b>	<b>0.71</b>	<b>0.75</b>								

<sup>a</sup>The maximum value in each column section is bold. The PSI-Blast alignment ADA-URE (marked by \*) has an *e*-value worse than  $1e-5$  and was excluded from the input to MaxFlow. Zeros in the PSI-Blast row mean that the alignment has an *e*-value worse than 1 and is not reported in the database search. The iterative profile search by PSI-Blast is asymmetrical depending on which protein in the pair is used as query, so reciprocal protein-protein hits lead to two (nonidentical) alignments. The MEME server limits the input to 60,000 characters, and the results are for a reduced representative set. No results were obtained from the T-Coffee program (out of memory) or the MatchBox server (no reply). None of the pairs is aligned by Blast at *e*-values  $< 1$ .

TABLE 4. HIGH SCORING SEGMENTS ARE MORE RELIABLE<sup>a</sup>

<i>Dots considered</i>	<i>Shift = 0</i>		<i>Shift = 0–5</i>	
	<i>Reliability</i>	<i>Coverage</i>	<i>Reliability</i>	<i>Coverage</i>
$M(s, t) > 0.0$	0.38	0.41	0.70	0.75
$M(s, t) > 0.1$	0.38	0.41	0.70	0.74
$M(s, t) > 0.2$	0.38	0.39	0.69	0.70
$M(s, t) > 0.4$	0.45	0.34	0.71	0.45
$M(s, t) > 0.6$	0.64	0.30	0.86	0.41
$M(s, t) > 0.8$	0.72	0.16	0.90	0.19
$M(s, t) = 1.0$	0.77	0.11	0.94	0.13

<sup>a</sup>Optimal alignments of the 10 pairs in the urease superfamily are evaluated at different thresholds of the MaxFlow score  $M(s, t)$ .

PSI-Blast (Schaffer *et al.*, 2001) identifies roughly twice as many correct residue pairs as does ClustalW or Dialign2, even though there are no direct PSI-Blast alignments linking PTE to any of the other structures. As a result, all tests involving PTE fail with PSI-Blast in Table 3. Coverage is improved and accuracy sustained on applying the MaxFlow procedure to a PSI-Blast input library. Not only does MaxFlow generate accurate alignments where PSI-Blast detects no statistically significant sequence similarity, but MaxFlow also extends the correct alignment (shift = 0) further than PSI-Blast in two cases (URE-DHO and CDA-URE). In some cases, a PSI-Blast alignment scores more highly in accuracy. However, our approach has the advantage that MaxFlow scores can be used as a predictor of reliable regions in the alignment. MaxFlow's coverage is higher than that of any competing method at their level of reliability (Table 4).

Despite highly variable sequences around the invariantly conserved functional residues, the path model of MaxFlow detects a strong clear signal derived from consistent alignments around the signature motif (Fig. 5). A sequence derived consensus alignment based on MaxFlow scores correctly identifies the first *H.H* motif in four out of five sequences (missed in ADA, see Fig. 5), the second *H* motif in all five, the third *H* motif in four out of five (missed in URE), and the *D* motif in all five sequences. It is particularly noteworthy that all motifs in PTE are correctly aligned despite only indirect PSI-Blast links in the trace graph and a “narrow” bridge near DHO in the protein adjacency matrix (see Fig. 2). The optimal alignment is well defined, as the optimal score was consistently about twice as high as the score of the second-best alignment (data not shown). We required that suboptimal alignments were nonoverlapping with any higher scoring alignments. Compared to PSI-Blast, MaxFlow achieved double coverage without compromising reliability.

Details of two more typical examples are discussed below. Both superfamilies have many pairs of known structures that are only indirect PSI-Blast neighbors.

### 6.2. DNA polymerase beta superfamily

DNA polymerase beta has a binding site at a compact  $\gamma\beta\beta$ -unit, while the rest of the structure lacks specific sequence motifs but has long insertions/deletions involving peripheral secondary structure elements. The signature motif is  $G.\{10, 11\}D.[DE].\{23, 85\}[DE]$ . Four structures were known in this superfamily. MaxFlow identifies the motif in the active site between all pairs (Table 5). The nominal reliability of PSI-Blast is higher than that of MaxFlow, because PSI-Blast detects only one pair which has 24% sequence identity. Interestingly, MaxFlow is able to extend the correct (shift = 0) alignment of this pair (TDT-POL), in addition to increasing coverage through the whole superfamily.

### 6.3. Actin superfamily

The ATP-binding domains of the actin superfamily share five subtle motifs where there is only one strictly invariant glycine residue. The motif defined by the structural alignment of eight proteins is  $D.G.\{96, 257\}G.\{20, 67\}G.\{88, 113\}G.\{26, 40\}G$  where the first four positions contact ATP and the fifth is a hinge point between two domains that undergo interdomain motion. This is an even more challenging

TABLE 5. ACCURACY OF SEQUENCE DERIVED ALIGNMENT IN THE DNA POLYMERASE BETA SUPERFAMILY<sup>a</sup>

Method	Shift	TP						Total TP	Rel. TP/P	Cov. TP/T
		POL-PAP	POL-TDT	POL-KAN	PAP-TDT	PAP-KAN	TDT-KAN			
Dali (T)	0	<b>135</b>	<b>307</b>	<b>107</b>	<b>134</b>	<b>98</b>	<b>107</b>	<b>888</b>	<b>1.00</b>	<b>1.00</b>
PSI-Blast	0	0	270	0	0	0	0	270	<b>0.87</b>	0.30
MaxFlow	0	<b>29</b>	<b>280</b>	<b>45</b>	<b>29</b>	<b>48</b>	<b>45</b>	<b>476</b>	0.61	<b>0.54</b>
PSI-Blast	0-5	0	<b>306</b>	0	0	0	0	306	<b>0.98</b>	0.34
MaxFlow	0-5	<b>55</b>	305	<b>56</b>	<b>55</b>	<b>54</b>	<b>56</b>	<b>581</b>	0.76	<b>0.65</b>

<sup>a</sup>Only the better of reciprocal PSI-Blast hits is reported.

TABLE 6. ACCURACY OF PAIRWISE PSI-BLAST ALIGNMENTS IN THE ACTIN SUPERFAMILY<sup>a</sup>

PSI-Blast (TP)	ACE	GLU	FTS	HXK	HSP	MRE	ACT	GLK
ACE	—	0	0	0	0	0	0	0
GLU	0	—	47	0	0	23	0	53
FTS	0	75	—	0	96	107	63	<b>47</b>
HXK	0	0	0	—	0	0	0	0
HSP	0	0	<b>185</b>	0	—	145	118	38
MRE	0	23	164	0	238	—	148	<b>47</b>
ACT	0	0	156	0	<b>240</b>	209	—	0
GLK	0	63	75	0	106	80	0	—

<sup>a</sup>Upper triangle: shift 0; lower triangle: shift 0-5. Values larger than the corresponding entry in Table 7 are bold. Only the better of reciprocal PSI-Blast hits is reported.

TABLE 7. ACCURACY OF PAIRWISE MAXFLOW ALIGNMENTS IN THE ACTIN SUPERFAMILY<sup>a</sup>

MaxFlow (TP)	ACE	GLU	FTS	HXK	HSP	MRE	ACT	GLK
ACE	—	<b>22</b>	<b>18</b>	<b>49</b>	<b>17</b>	<b>17</b>	<b>19</b>	<b>44</b>
GLU	<b>64</b>	—	<b>69</b>	<b>44</b>	<b>47</b>	<b>67</b>	<b>44</b>	<b>66</b>
FTS	<b>54</b>	<b>135</b>	—	<b>21</b>	<b>101</b>	<b>133</b>	<b>101</b>	43
HXK	<b>114</b>	<b>81</b>	<b>90</b>	—	<b>16</b>	<b>27</b>	<b>16</b>	<b>53</b>
HSP	<b>71</b>	<b>118</b>	184	<b>92</b>	—	145	<b>119</b>	<b>45</b>
MRE	<b>54</b>	<b>102</b>	<b>197</b>	<b>81</b>	<b>263</b>	—	<b>157</b>	44
ACT	<b>71</b>	<b>82</b>	<b>169</b>	<b>73</b>	211	<b>212</b>	—	<b>46</b>
GLK	<b>85</b>	<b>171</b>	<b>116</b>	<b>99</b>	<b>156</b>	<b>107</b>	<b>120</b>	—

<sup>a</sup>Upper triangle: shift 0; lower triangle: shift 0-5. Values larger than the corresponding entry in Table 6 are bold.

case than the urease superfamily. Again, MaxFlow increases the coverage of structurally equivalent positions while maintaining comparable reliability to PSI-Blast (Tables 6-7). ACE and HXK have no direct PSI-Blast links to the other known structures, but HXK in particular is quite decently aligned by MaxFlow.

MaxFlow recovers the motif with quite remarkable specificity, considering that the motifs are difficult to align precisely, as there are long insertions/deletions between the motif positions and the invariant residues are glycines in a general hydrophobic environment at the end of beta-strands. As a result, the input library generated by PSI-Blast contains many shifted alignments. Of five motif positions in 28 pairwise comparisons of the eight known structures, MaxFlow recovers 30% correct matches at shift = 0 and 61% correct matches at shift = 0-5.

TABLE 8. ACCURACY OF PAIRWISE MAXFLOW ALIGNMENTS IN THE LARGE BENCHMARK

<i>Set</i>	<i>Method</i>	<i>Shift</i>	<i>Coverage</i>	<i>Reliability</i>
Direct PSI-Blast neighbors: 258 test pairs	MaxFlow	0	0.52	0.62
	PSI-Blast	0	0.51	0.59
	MaxFlow	0–5	0.64	0.77
	PSI-Blast	0–5	0.65	0.74
Direct to fourth PSI-Blast neighbors: 420 test pairs	MaxFlow	0	0.46	0.58
	PSI-Blast	0	0.34	0.59
	MaxFlow	0–5	0.58	0.74
	PSI-Blast	0–5	0.44	0.74

#### 6.4. Benchmarking against the Protein Data Bank

The large benchmark test indicates that the improvement over PSI-Blast holds in general. The results are summarized in Table 8 (see <http://www.bioinfo.biocenter.helsinki.fi:8080/MaxFlow> for detailed listing). MaxFlow aligns also indirect PSI-Blast neighbors, which increases the coverage by MaxFlow while the reliability remains high. For the direct PSI-Blast neighbors, MaxFlow produces a slight improvement of reliability (about three percentage-points). Indirect PSI-Blast neighbors are aligned by MaxFlow at almost the same accuracy as the direct neighbors are aligned by PSI-Blast.

## 7. DISCUSSION

Optimal sequence alignment is at the core of two central questions in bioinformatics: homology detection and alignment accuracy. Alignment score statistics are used to discriminate related sequences from unrelated ones. Given a scoring function, one can always determine an optimal alignment between two sequences or between a sequence and a profile model. The optimum of the scoring system may or may not coincide with the structural alignment, the standard of truth. Position-specific profile or HMM models of sequence evolution are statistically rigorous, but have a limited radius of detection and reliable alignment (cf. Heger and Holm, 2003a). MaxFlow yields high quality alignments between very distant homologues based on transitive alignments. MaxFlow uses a heuristic score to evaluate transitive alignments in terms of the weakest link, but this leads to a simple algorithmic solution.

The requirements for MaxFlow are that the aligned proteins belong to a connected set. Obviously, a correct signal must be present in the input library of pairwise alignments as MaxFlow does not create any new edges in the alignment trace graph. If the input library contains random noise, it should cancel out on average, and this is how MaxFlow (and consensus methods in general) achieve an improvement. The improvement over PSI-Blast is most marked between remote homologues, where the input library has higher variance. The sequence alignment score of Blast *et al.* can be converted to bits that reflect information content; at very low scores, the optimal alignment trace becomes indistinguishable from random alignments (Altschul, 1991).

The present prototype implementation runs comfortably on input sets of up to 500 representative sequences. MaxFlow has an edge over PSI-Blast at long sequence distances, especially between indirect PSI-Blast neighbors. Higher consistency scores empirically correlate with higher reliability of alignment. High scoring segments can also define useful anchor points for alignment refinement in 3D model building by homology (Bork *et al.*, 1995; Flohil *et al.*, 2002). Accurate multiple alignments provide input to methods for the computational identification of functional sites (Madabushi *et al.*, 2002; Casari *et al.*, 1995), although the requirement of explicit alignment can be relaxed (Heger and Holm, 2003b).

## ACKNOWLEDGMENT

We thank Harshad Joshi for the benchmarking data.

## REFERENCES

- Altschul, S.F. 1991. Amino acid matrices from an information theoretic perspective. *J. Mol. Biol.* 219, 555–565.
- Bork, P., Holm, L., Koonin, E., and Sander, C. 1995. The cytidylyltransferase superfamily: Identification of the nucleotide-binding site and fold prediction. *Proteins* 22, 259–266.
- Casari, G., Sander, C., and Valencia, A. 1995. A method to predict functional residues in proteins. *Nat. Struct. Biol.* 2, 171–178.
- Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M., and Holm, L. 2001. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucl. Acids Res.* 29, 55–57.
- Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* 14, 755–763.
- Flohil, J.A., Vriend, G., and Berendsen, H.J.C. 2002. Completion and refinement of 3-D homology models with restricted molecular dynamics: Application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis. *Proteins* 48, 593–604.
- Grundy, W.N., Bailey, T.L., Elkan, C.P., and Baker, M.E. 1997. Meta-MEME: Motif-based hidden Markov models of protein families. *CABIOS* 5, 211–221.
- Heger, A., and Holm, L. 2003a. More for less in structural genomics. *J. Struct. Funct. Genomics*. In press.
- Heger, A., and Holm, L. 2003b. Sensitive pattern discovery with “fuzzy” alignments of distantly related proteins. *ISMB'03*. In press.
- Holm, L., and Sander, C. 1997. An evolutionary treasure: Unification of a broad set of amidohydrolases related to urease. *Proteins* 28, 72–82.
- Kececioğlu, J. 1993. The maximum weight trace problem in multiple sequence alignment. *Proc. 4th Symposium on Combinatorial Pattern Matching*, No. 684 in *Lect. Notes Comput. Sci.* 106–119.
- Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* 262, 208–214.
- Lindahl, E., and Elofsson, A. 2000. Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* 295, 613–625.
- Madabushi, S., Yao, H., Marsh, M., Kristensen, D.M., Philippi, A., Sowa, M.E., and Lichtarge, O. 2002. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* 316, 139–154.
- Morgenstern, B. 1999. DIALIGN2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* 15, 211–218.
- Notredame, C. 2002. Recent progress in multiple sequence alignment: A survey. *Pharmacogenomics* 3, 131–144.
- Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302, 205–217.
- Sander, C., and Schneider, R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 19, 56–68.
- Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., and Altschul, S.F. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucl. Acids Res.* 29, 2994–3005.
- Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* 22, 4673–4680.
- Vingron, M., and Argos, P. 1991. Motif recognition and alignment for many sequences by comparison of dot-matrices. *J. Mol. Biol.* 218, 33–43.

Address correspondence to:  
Andreas Heger  
Institute of Biotechnology  
P.O. Box 56  
00014 University of Helsinki  
Helsinki, Finland

E-mail: Andreas.Heger@Helsinki.fi