Database search based on Bayesian alignment

Jun Zhu, Roland Lüthy
Department of Computational Biology, MS-29-2-A, Amgen, Inc.
One Amgen CenterDrive, Thousand Oaks, CA 91320
E-mail: junz, rluethy@amgen.com, Tel: 805-447-1765, 8185, Fax: 805-499-9955

Charles E. Lawrence
Wadsworth Center for Laboratories & Research, New York State Department of Health
P.O. Box 509, Empire State Plaza, Albany, NY 12201
E-mail: lawrence@wadsworth.org, Tel: 518-473-3382, Fax: 518-473-2900

Abstract

The size of protein sequence database is getting larger each day. One common challenge is to predict protein structures or functions of the sequences in databases. It is easy when a sequence shares direct similarity to a well-characterized protein. If there is no direct similarity, we have to rely on a third sequence or a model as intermediate to link two proteins together. We developed a new model based method, called Bayesian search, as a means to connect two distantly related proteins. We compared this Bayesian search model with pairwise and multiple sequence comparison methods on structural databases using structural similarity as the criteria for relationship. The results show that the Bayesian search can link more distantly related sequence pairs than other methods, collectively and consistently over large protein families. If each query made one error on average against SCOP database PDB40D-B, Bayesian search found 36.5% of related pairs, PSI-Blast found 32.6%, and Smith-Waterman method found 25%. Examples are presented to show that the alignments predicted by the Bayesian search agree well with structural alignments. Also false positives found by Bayesian search at low cutoff values are analyzed.

Introduction

Since the number of protein (or DNA) sequences is increasing much faster than the number of structures. Protein structure prediction is becoming even more important. Homology modelling is the most promising way to predict a protein structure. The homology of sequences can be found by single sequence comparison methods, such as BLASTP (Altschul *et al.* 1990), FASTA (Pearson 1988), Smith-Waterman (S-W) (Smith & Waterman 1981). The sequences of related proteins can be so divergent that their relationship can not be detected reliably by these programs. However, two distantly related sequences may be both closer to a third sequence. The third sequence is acting as an intermediate sequence which can connect the two sequences together. The straight forward implementation of this idea is the intermediate sequence method (IS) (Park *et al.* 1997). Most multiple sequence model methods, such as profile (Gribskov, McLachlan, & Eisenberg 1987; Lüthy, Xenarios, & Bucher 1994; Thompson, Higgins, & Gibson 1994b), PROBE (Neuwald *et al.* 1997) or Gibbs sampler (Lawrence *et al.* 1993), Hidden Markov models (HMM) (Krogh *et al.* 1994; Baldi *et al.* 1994; Eddy 1995; Eddy, Mitchison, & Durbin 1995), and Position-Specific-Iterative BLAST (PSI-BLAST) (Altschul *et al.* 1997), are also based on this idea. In these cases, a set of diverged sequences, which share common features, are acting as intermediate sequences. The common features are expressed as a motif or profile. There are some limitations for the popular multiple sequence model methods, e.g., HMM needs relative large number of sequences to start with and require extensive computational resources; Gibbs sampler requires that all motifs must exist and exist intactly. We present here a new profile-like model, called Bayesian search, which uses Bayesian statistical inference to combine the data from a collection of intermediate sequences. The distinguishing features of the Bayesian search model method are: (1) the model is not derived from the optimal alignment of sequences; (2) sequences can have only some of the motifs or part of a motif; (3) each residue in each sequence has a different weight of contribution to the model based on marginal posterior probability that the residue aligns with the query sequence; (4) each position in the query sequence (column in model) has a different weight depending on the probability that the position is conserved.

Methods

Most multiple sequence methods use a two-step process: finding potential related sequences, and capturing information from them. The first step is to identify intermediate sequence candidates. Pairwise methods, such as BLAST, FASTA, Smith-Waterman, transitive BLAST (Neuwald *et al.* 1997), are often used for this step. It is common that liberal definition of relatedness is applied. The second step is to decide what information to use from the candidate sequences. Not all of the residues in the candidate sequences will be related to all the residues in the query sequence. In fact, since candidates may have been recruited with liberal rules, perhaps none of a particular candidate sequence is related to the query. IS method directly uses all the information in aligned regions of candidate sequences. PSI-BLAST aligns candidate sequences with query sequences using gapped BLAST, then estimates the probability of residue i at each position. Similarly, HMM builds alignments using BLAST, then estimates a model using a recursive relationship in an iterative fashion and applying transition regularizer and Dirichlet mixture regularizer to alignments. While, PROBE uses Gibbs sampler to build alignments, then motif models are created.

Bayesian search model construction is also a two-step process. Its first step is similar to IS, PSI-BLAST and PROBE. The second step differs. Bayesian search is based on alignments generated by Bayes aligner (Zhu, Liu, & Lawrence 1998), which has the following features: (1) it does not use optimal alignments, instead it finds the posterior probability distribution for the alignment of residues in the query sequence with residues in a candidate sequence; (2) it is a super local alignment, which consists of several local alignments (3) it does not use gap penalties, instead it uses the number of local alignment regions to constrain alignments; (4) it uses full series of scoring matrices rather than a single one; (5) it identifies varying levels of conservation over the sequence.

After a multiple sequence model is constructed, it can be compared with other sequences. For example, PSI-BLAST uses gapped BLAST and HMM use the Smith-Waterman algorithm. Here the multiple sequence models are compared with sequences using the Bayes aligner.

The following is the detail steps of model construction and comparison of models and sequences.

Constructing Bayesian search models using Bayes sequence aligner

As most multiple sequence methods do, Bayesian search model construction starts by collecting sequences similar to the query sequence in the database. The similar sequence recruiting phase uses transitive BLAST (Neuwald *et al.* 1997) as following: the query sequence is searched against a non-redundant database (NR). After the pairwise search, the sequences related to the query are used as queries to search for additional similar sequences. The process is repeated recursively until no new related sequences can be found. Then, the resulting set of similar sequences is purged at a cutoff score to remove closely related sequences.

Each sequence $R^{(n)}$ in the recruited set can be aligned with the query sequence Q by Bayes sequence aligner. Their alignment can be characterized by an indicator matrix $A_{i,j}$. If Q_i and $R_j^{(n)}$ are aligned, $A_{i,j}$=1, otherwise $A_{i,j}$=0. The probability of the alignment is represented by marginal posterior alignment distribution $P(A_{i,j} = 1|Q, R^{(n)})$. The model $M(i,a)$, the probability of amino acid a at position i, is

$$M(i,a) = \frac{pseudo * \theta_a + \sum_{n,R_j^{(n)}=a} P(A_{i,j} = 1|Q, R^{(n)})}{\sum_a pseudo * \theta_a + \sum_n P(A_{i,j} = 1|Q, R^{(n)})}, \quad (1)$$

where *pseudo* is pseudo count, which is proportional to the square root of the number of sequences in the data set, $\sqrt{N}$; θ_a is the prior probability of observing amino acid a, estimating from included sequences.

There are two levels of weight when constructing the model M: (1) every individual residue in each sequence is weighted differently depending on $P(A_{i,j} = 1, R_j^{(n)} = a|Q, R^{(n)})$, the probability that amino acid a at position j in sequence $R^{(n)}$ is aligned to position i of query sequence. (2) All residues in a column of the model (a position in query sequence) are weighted differently from residues in other columns, depending on $\sum_n P(A_{i,j} = 1|Q, R^{(n)})$, the confidence that the position in query sequence is conserved. When $\sum_n P(A_{i,j} = 1|Q, R^{(n)})$ is high, the prior information (pseudo counts) will be weighted down. And *vice versa.*

Bayesian search model is similar to general profile or motif models. The psuedo count used here is similar to Gibbs sampler (Lawrence *et al.* 1993), has no data-dependency. Both profile model (Lüthy, Xenarios, & Bucher 1994) and PSI-Blast model (Altschul *et al.* 1997) use data-dependent psuedo counts. One key difference of Bayesian search model to others is that the ratio of psuedocounted and observed frequencies is not fixed, but varies from column to column as mentioned above. $\sum_{n,R_j^{(n)}=a} P(A_{i,j} = 1|Q, R^{(n)})$ has the similar meaning as the observed frequency used in Gibbs sampler, profile and PSI-Blast models. The difference is that an amimo acid a in a column i will be counted as 1 in Gibbs sampler, profile and PSI-Blast models, but will be counted according to $P(A_{i,j} = 1, R_j^{(n)} = a|Q, R^{(n)})$.

We assume that each sequence contributes equally to the model because closely related sequences are purged. Neuwald *et al.* (1997) show that it has a similar effect as the weighting scheme described by Henikoff and Henikoff (1994).

Comparing the Bayesian models with sequences

Our goal is to connect two distantly related sequences using the Bayesian search model. Thus, the pairwise sequence comparison is converted to a model-sequence comparison. A Bayesian search model is a set of position specific probabilities of amino acid a at position i. All pairwise sequence alignment methods can be converted into model-sequence comparison methods by substituting general rationship matrices, such as PAM (Schwartz & Dayhoff 1978) or BLOSUM (Henikoff & Henikoff 1992) matrices, with a multiple sequence

Error	Coverage (%)				
(EPQ)	S-W	P-B(2/0.01)	P-B(20/0.01)	P-B(20/0.0005)	Bayesian
0.01	18.01	20.34	12.61	14.48	14.23*
0.1	20.89	23.65	25.12	24.90	24.46
0.15	21.58	24.43	26.78	26.43	26.91
0.25	22.18	25.14	28.93	27.81	29.70
0.5	23.37	26.21	30.97	29.58	33.54
0.75	24.45	26.95	32.00	30.30	35.33
1	25.01	27.28	32.56	30.79	36.49
1.5	25.91	28.11	33.41	31.40	37.70
2.0	26.81	28.51	33.95	31.91	38.54

Table 1: Coverage (true positive/total positive, true positive excludes query sequence itself) at different error per query (EPQ) when searching against PDB40D-B database and using SCOP as criteria to determine relationship. S-W: Smith-Waterman method; P-B(2/0.01):PSI-Blast with 2 iteration and e-value (e_m) for inclusion in multipass model set to 0.01 (default);P-B(20/0.01): PSI-Blast with 20 iteration and e_m set to 0.01 (default); P-B(20/0.0005): PSI-Blast with 20 iterations and e_m set to 0.0005, which is suggested by Park *et al.* (1998). Bayesian: Bayesian search. Bayesian search performed better than S-W and PSI-Blast with EPQ >0.15. (* This coverage is at 0.0128 EPQ, which is lowest error rate we can reach because there are 17 false positives with Bayesian evidence equal to 0.)

model. Here we convert the Bayes sequence aligner to a model-sequence comparison method.

Considering a Bayesian search model M, a sequence R, and an alignment A, similar to the definition in Zhu *et al.* (1998), the likelihood of the model and sequence conditioned on the alignment is

$$logP(M_i, R_j|\theta, A) = \theta_{R_j} + A_{i,j}logM(i, R_j); \qquad (2)$$

and

$$logP(M, R|\theta, A) = \sum_j \theta_j + \sum_{i,j} logM(i, R_j) \qquad (3)$$

Assuming that θ, the marginal probability of an amino acid, is known and fixed, means we can ignore it. Bayes sequence aligner do not use gap penalties. Instead, it uses the number of of local alignment regions to constrain alignments, similar to Sankoff (Sankoff 1972)'s approach. The alignments are catagorized according to number of aligned blocks. A k-block alignment is an alignment which contains k aligned blocks (or, equivalently, k+1 gaps). Without *a priori* information, we employ uninformed prior for all priors setting. $P(k)$ (prior probability of k-block alignments) $P(A|k)$ (prior probability of a k-block alignment A) are the same as in Zhu *et al.*(1998). Using those priors, we can derive the joint probability of model M, sequence R, and a k-block alignment A,

$$P(M, R, A, k) = P(M, R|A, k)P(A|k)P(k). \qquad (4)$$

And the posterior probability of a k-block alignment is

$$P(k|M, R) = \frac{\sum_A P(M, R|A, k)P(A|k)P(k)}{\sum_k \sum_A P(M, R|A, k)P(A|k)P(k)} \qquad (5)$$

The Bayesian evidence that model M and sequence R are not related is $1 - sup_{P(k)}\{P(k > 0|M, R)\}$, which has similar meaning as a p-value and is calculated in the same way as in Zhu *et al.* (1998). If the model M and sequence R are related, we can infer that the sequence R^0 from which the model M is built and this sequence R are related through the connection by the model M. The alignment of R^0 and R can be expressed also through the model as marginal posterior alignment distribution, $P(A_{i,j} = 1|M, R)$. The higher the value, the more confident we are that the sites in the two proteins are related.

Results

To get a fair comparison of sequence alignment methods, we need to define true relationship based on criteria independent from sequence information. We applied the sequence alignment methods to structural database, and used structural similarity as the criteria for relatedness of two sequences. We used the hierarchical classification of structures in SCOP as structure similarity criteria.

Testing on structural databases

We compared this Bayesian search with the Smith-Waterman method, which is the best pairwise alignment method (Brenner, Chothia, & Hubbard 1998). We also compared the Bayesian search with another multiple sequence method: PSI-BLAST. We applied the S-W method, PSI-Blast and the Bayesian search to the structure domain database pdb40D-B in SCOP version 1.35 (Brenner, Chothia, & Hubbard 1998). Positive relationships are defined by the SCOP classification. This approach is similar to Brenner et al. (1998)'s. For multiple sequence alignment methods, after we build an initial model, we can use the model to recruit more similar sequences in the database and refine the model iteratively. However, the most critical step is to build the initial model. Subsequent iterations and refinements give only small improvement. Here we used only one

familiy	family size	Bayesian	PSI-Blast
Immunoglobulin (2.1.1)	41	481	215
NAD(P)-binding Rossmann-fold domains (3.18.1)	27	123	132
Glycosyltransferases (3.1.1)	23	92	76
Viral coat and capsid proteins (2.8.1)	22	39	40
Trypsin-like serine proteases (2.29.1)	20	260	192
FAD/NAD(P)-binding domain (3.4.1)	19	108	86
Cupredoxins (2.5.1)	17	72	44
P-loop containing nucleotide triphosphate hydrolases (3.24.1)	16	71	26
Globin-like (1.1.1)	16	157	134
Membrane all-alpha (6.4.1)	15	3	3
Acid proteases (2.32.1)	14	90	81
Homeodomain-like (1.4.1)	14	51	32
EF-hand (1.31.1)	14	143	117
alpha/beta-Hydrolases (3.48.1)	14	40	22
Classic zinc finger, C2H2 (7.30.1)	13	139	100
Periplasmic binding protein-like II (3.68.1)	12	18	14
Cytochrome c (1.3.1)	12	108	54
Thioredoxin-like (3.30.1)	11	35	22
EGF/Laminin (7.10.1)	11	75	63
ConA-like lectins/glucanases (2.18.1)	11	20	20
4-helical cytokines (1.22.1)	11	2	0
Bacterial enterotoxins (2.24.1)	10	3	3
Fibronectin type III (2.1.2)	10	40	26
Lipocalins (2.39.1)	10	37	35
Total	383	2207	1537

Table 2: Number of pairs of true relationship found by Bayesian search model and PSI-Blast at 1 EPQ for SCOP families with 10 or more members in pdb40D-B. The family is identified by family name and ID in SCOP. The number of total pairs of relationship is n*(n-1), where n is the family size (number of sequences in pdb40D-B belonging to that family). The cutoff value is family-specific, i.e., cutoff values for different families are different, and for an individual family one false positive is found for each query on average. Only the number of true relationships found is displayed here. The number of false positives found is equal to the family size.

iteration for Bayesian search. For PSI-Blast we used both a single iteration and multiple iterations.

The parameters of each program were set as following:

SSEARCH in the FASTA3 package was used as Smith-Waterman algorithm implementation. The parameters used were scoring matrix BLOSUM45 with gap penalty -12/-1, e-value as selecting criteria. This combination is the best for detecting remote similarity as suggested by Brenner *et al.* (1998).

The PSI-Blast model was trained against the NR database. All parameters were set to default values except the number of iterations and e-value (e_m) for inclusion in multipass model. We made three combinations: (1) number of iterations set to 2 and e_m set to 0.01 (default); (2) number of iterations set to 20 and e_m set to 0.0005, which suggested by Park *et al.* (1998); (3) 20 iterations and e_m set to 0.01. The resulting model was saved, and was later used to search the pdb40D-B database. The e-value was used as the criteria to determine similarity between a model and a sequence.

A Bayesian search model for each sequence pdb40D-B was constructed as following: First, for each sequence we searched against the NR database using transitive BLAST (Neuwald *et al.* 1997) with e-value equal to 1.0. Highly similar sequences were purged using a BLAST score of 150 as the cutoff. Then, the set of sequences was aligned with the query sequence and the Bayesian search model was built as described in the Method Section. The relation matrices used by Bayes aligner were the BLOSUM series (Henikoff & Henikoff 1992), 30, 35, 40, 45, 55, 62, 80, 100. The number of blocks k included in a model is defined by posterior probability $P(K >= k|R, Q) > 0.8$. If a sequence has no aligned blocks with the query, the sequence is removed from the set of sequences similar to the query. Finally, this model is compared with all the sequences in pdb40D-B. Bayesian evidence is used to determine relationships.

There were 1323 domains in PDB40D-B, 9044 pairs of distant relationships, and a total of 1,749,006 pairs in all against all comparison. We used the same error measurement, error per query (EPQ), as Brenner *et al.*

```
a                       10        20        30        40        50        60        70        80
                         |         |         |         |         |         |         |         |
cH-p21          MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLC
SW:RAS_GEOCY    MTEYKIVIVGGGLVGKSALTLQLVQNHYIDEYDPTVEDSXRREVSIDDQTCLLNILDTAGQQHSNAQSXXXXXXSTVFVC
GP:RMRRASX23_1  XXXXXXXXXXXXXXXXXXXXXXXXXSYFVTDYDPTIEDSYTKQCVIDDRPARLDIDTAGLQEEFGAMREQYMRTGEGFLL
PIR:A31798      XXXXXXXXXXXXXXXXSALTIQLIQNHFVDEYDPXIEDSYRXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
GP:HSNRASE1_1   MTEYKLVVVGAGGVGKSALTIQLIQNHFVXXXXXXXXXXXXXQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFXX
PIR:I48307      XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
SW:IF2_STIAU    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXVDHGKTSLLTFLDTPGHEAFTSMRARGANVTDIVIL
GP:CEC08F8_8    MKEYKIVVLGNGGVGKSALTLQYVQGIFVHTYDATIEDSYRKLSKVDAENARLEILDTAGQEQFTGMRETYYRTAQGFVL
SW:YCR7_YEAST   NFQRKIALIGARNVGKTTLTVRFVESRFVESYYPTIENEFTRIIPYKSHDCTLEILDTAGQDEVSLLNIKSLTGVRGIML
GP:CELC54A12_4  TSDYRVAVFGAGGVGKSSITQRFVKGTFNENYVPTIEDTYRQVISCNKNVCTLQITDTTGSHQFPAMQRLSISKGNAFIL
GP:HSPAWN1_1    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXGREEYSAMRDQYMRTGEGFLC
GP:CELC14A11_7  EERYRLVVLGSAKVGKTNIIRRYLYNEFSSKYKETIEDLHSREFRIQGVPLPLDILDTNFXXNFPDMRRLSIASASAFLL
SW:IF2C_CYACA   LRAPIVAVLGHVNHGKTSLIEKLIKNDLTKAETGHITQIGAYEFIIGPKDKKIILLDTPGHEAFESIRQRVLKISDIILL
GP:MMU91601_1   EALYRVVLLGDPGVGKTSLASLFAEKQDRDPHEXQLGGVYERTLSVDGEDTTLVVMDTWEAEKLSWCQESCLQAGSAYVI
SW:RAB8_RAT     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXITTAYYRGAMGIML
PIR:S75088      XXMLRAGIVGLPNVGKSTLFNALVANAKAEAANFGSVDPVRDVEVIDLELVLADLAQVEKRLERSRKQARGNKXXXXXXI
GP:CELC56E6_2   KHKAKVVVLGDSGVGKTSIIYRHRYGAHYRPVNATIGASFXXXXXXREDVVRLQVWDTAGQERFRCMVPMYMRNADAALI
GP:S64261_1     MTEYKLVVVGAGGVGKSALTIQLIQNHFVDDYDPTLEXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
PIR:A46551      XXXXKLVVVGDGGVGKSALTIQFFQKLFVVDYDPTIEDSYLQYTEVDSEWCMLDXXXXXXXXXXXXXXXXXXXXXXXXXX
GP:CED1081_3    QNKVTVAVLGAERVGKSAMVSQFLWHKFVEDYRPTVEEFNWEYEIEEGRVLMVQIIDSSGSRDFIGMKNLYIGTADAFLV
GP:TBU18326_1   XXXXXXXXXXSASAGKSKLVERFLMQRXXXXXXXXXXXXXXDFVTEDDEAIDVDIWDTAGQXXXXXXXXXXXXXXXXXXX
GP:PTU03620_1   XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXTIGIEFATKTVTLQDGGKIQVQIWDTAGQERYRAITTNHFRGAGGALL
GP:LFU23469_1   XXXXXXXXXXXXXXXXXXXXXQFIQSYFVTDYDPTIEDSYTKQCQIDSVVARLDILXXXXXXXXXXXXXXXXXXXXXXXX
SW:RB28_HUMAN   DRQLKIVVLGDGTSGKTSLTTCFAQETFGKQYKQTIGLDFXXXXXXXNLNVTLQIWDIGGQTIGGKMLDKYIYGAQGVLL
SW:RB19_MOUSE   DYLFKVILIGDSNVGKTCVVQHFKSGVYSESQQNTIVDFTVRSLEIDGKKVKMQVWDTAGQEXXXXXXXXXXXXXXXXXX
SW:RAB6_MOUSE   LRKFKLVFLGEQSVGKTSLITRFMYDSFDNTYQATIGDFLSKTMYLEDRTIRLQLWDTAGQEXXXXXXXXXXXXXXXXXX
```

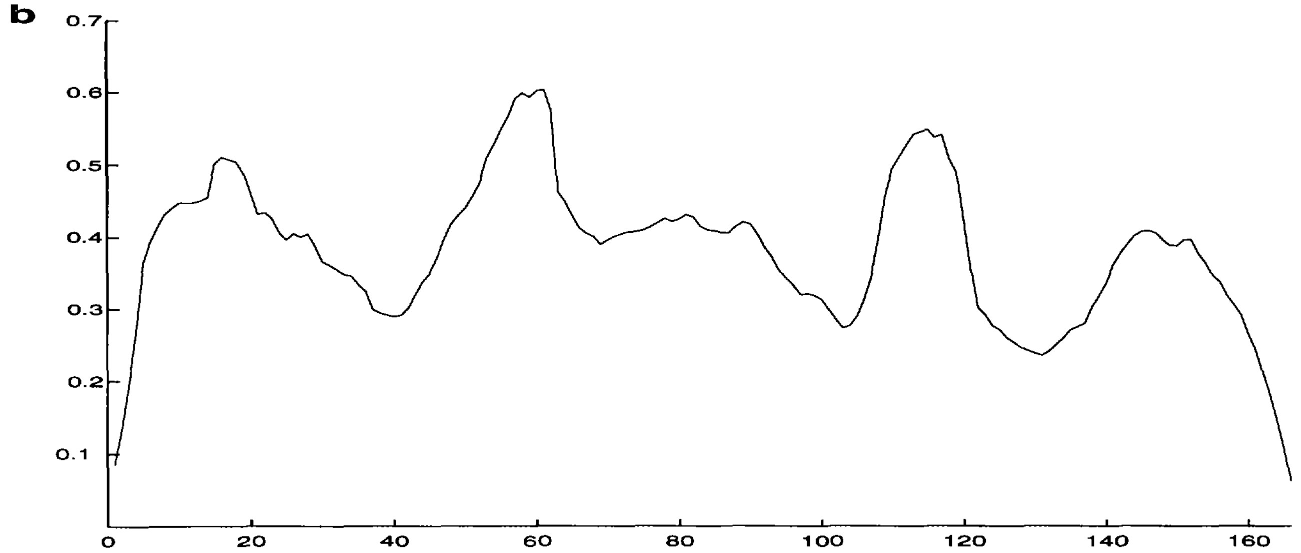

Figure 1: Bayesian model for cH-p21 Ras protein (in PDB, 5P21). 1a. Alignments used for building model for cH-p21. Only partial alignments (first 80 aa, and first 25 sequences) are presented here because of space limitation. The first row is the query sequence (cH-p21 Ras), the following rows are sequences of similar proteins recruited by transitive BLAST. The parts of sequences that are not aligned with the query sequence are marked as X. 1b. The probability that a site in cH-p21 is conserved, $\sum_n P(A_{i,j} = 1|Q, R^{(n)})/N$. X-axes is the amino acid position of cH-p21; Y-axes is the probability of conservation. The four peaks corresponds to 10-17, phosphate binding site; 57-61, Mg^{2+} binding site; 116-119 and 145-147, guanine nucleotide binding sites.

(1998) used. The coverages at different error rates are calculated, and listed in Table 1. At 1 EPQ, both PSI-Blast and Bayesian search perform better than S-W method.

To see how PSI-Blast and the Bayesian model method perform on a specific family at 1 EPQ error rate, we examined the families with 10 or more members in SCOP. The numbers of true and false positives found for each family are counted (result not shown). For most of families, Bayesian search found less false positives and more true positives. As they find different number of false positives, it is hard to tell how well each method really performs on a specific family. To ensure that both methods have the same error rate for a a

```
  1 MTEYKLVVVGAGGVGKSALTIQLIQ   25
            .....::::::
         ...::::::::::::
       .::::::::::::::::
      :::::::::::::::::::
    .::::::::::::::::::::::..
  1 KPHVNVGTIGHVDHGKTTLTAAITT   25

 52 LLDILDTAGQEEYSA   66
           .::.
        ::::::::
      ..:::::::::..
 45 HYAHVDCPGHADYVK   59

 71 YMRTGEGFLCVFAINNT   87
          .::::....
         ::::::::::.
         :::::::::::
        .:::::::::::
     ..::::::::::::::
 64 GAAQMDGAILVVAATDG   80

 92 DIHQYREQIKRVKDSD  107
        .::::.
      .::::::::..
     :::::::::::::..
    .:::::::::::::::
 81 PMPQTREHILLGRQVG   96

111 MVLVGNKCDLAARTVESRQAQDLARS  136
         ::::
     ...::::::.
    ::::::::::::.
    ::::::::::::::.
    :::::::::::::::::::::::::.
100 IIVFLNKCDMVDDEELLELVEMEVRE  125

140 PYIETSAKTR  149
         ..
    ::::::::.
138 PIVRGSALKA  147
```

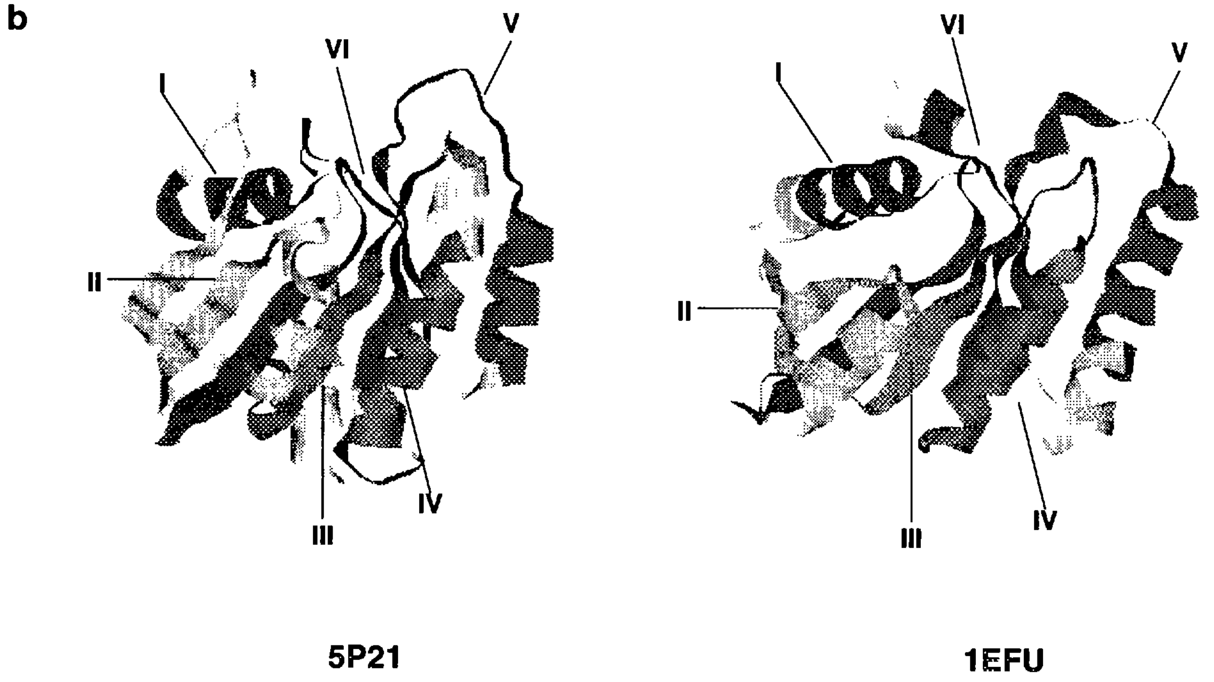

Figure 2: The alignment generated by the Bayesian model for cH-p21 ras protein (PDB 5p21) and elongation factor Tu (EF-Tu), the N-terminal (G) domain (PDB 1EFU). 2a. Aligned blocks found by the Bayesian model. The dots indicate the importance of aligned pairs. 2b. Structures of cH-P21 and EF-Tu.

protein family, we used a family-specific cutoff value at which each family had an error rate of 1 EPQ, instead of the universal cutoff used in all the above calculations. The number of detected positives is shown in Table 2. Some models in a family are more error prone than others, so we also applied sequence-specific cutoffs at which each sequence makes exactly one error. The result (not shown) is similar to the one in Table 2. It is clear that Bayesian search performs better on most large protein families.

The running time of building a Bayesian search model is comparable with that of multiple iteration PSI-

BLAST, each taking several minutes to build a model. While search a model against a database, PSI-Blast is faster than Bayesian search.

Examples – relationship found only by Bayesian search

The test on the SCOP database pdb40D-B, shown in Table 2, shows that the Bayesian search finds more structural neighbors than PSI-Blast for most of the families, except NAD(P)-binding Rossmann-fold domains, viral coat and capsid proteins. The alignments found by the Bayesian search agree well with structural alignments. We will examine one of these in detail.

All nucleotide triphosphate hydrolases (3.24.1 in SCOP) contain a P-loop to bind phosphate. Both structure and function are conserved among proteins in the family. The coverage of the Bayesian search for this family is 29.6% (71/16*15), much better than PSI-BLast's performance, 10.8%. Human Ras protein cH-p21 (PDB 5P21) (Pai *et al.* 1990), which hydrolyses GTP to GDP, is a member of this family, is used as the query. The alignment of the first 80 residues used to build the Bayesian search model for cH-p21 is shown in Fig 1a. Only regions conserved with cH-p21 are aligned, other parts are ignored by the Bayesian aligner. $\sum_n P(A_{i,j} = 1|Q, R^{(n)})$, which indicates the degree of conservation along cH-p21, is plotted as Fig 1b. There are four peaks in Fig. 1b, which represent four predicted conserved regions. The predicted conserved regions agree well with functional/structural studies (Pai *et al.* 1990): 10-17, phosphate binding site; 57-61, Mg^{2+} binding site; 116-119 and 145-147, guanine nucleotide binding sites. There are three other function sites of cH-p21 (Boriack-Sjodin *et al.* 1998), switch I, switch II and α3-L7 region, for interacting with guanine exchange factor or guanine activating factor. The functions of switch I (25-40) and α3-L7 region are different in different subfamilies of GTPase (Boriack-Sjodin *et al.* 1998; Day, Mosteller, & Broek 1998). The sequences in those regions are only conserved in subfamilies, but not among all GTPase family . While function of switch II (57-75), conformational changes to turn on/off Mg^{2+} binding site, are similar among GTPase. They all form a hydropillic helix. However, only residue Glu62 is conserved among GTPase family. Other sequences in the region are only conserved in sub-families (Day et al. 1998). Fig 1b shows that residue 62 is very conserved.

Bayesian search linked cH-p21 to another 8 out of 15 triphosphate hydrolases in pdb40D-B, while PSI-BLAST only found 3 pairs of relationship. The relationship to another GTPase, elongation factor Tu (EF-Tu) from E. Coli (PDB 1EFU, region 1-174) (Kawashima *et al.* 1996), which is only found by Bayesian search, has the Bayesian evidence that the two are not related as 2.64304e-07, while the cutoff value for this family at 1 EPQ is 0.0627. The alignment of these two proteins includes all above functional sites, shown as Fig. 2a. Among those sites, phosphate and guanine binding sites are the most conserved. The alignment created by Bayesian search agrees well with the structural alignment, shown as Fig. 2b.

Examples – false relationship found by Bayesian search

If there is no trace of sequence conservation among members in a protein family, Bayesian search will identify common secondary structure elements. It is often the case that it will also pick similar structure elements in proteins from other families. It is hard to find relationships reliably among those families by sequence similarity search. Both Bayesian model and PSI-Blast perform poorly on these cases.

At low error rate (< 0.15 EPQ), Bayesian search performed worse than PSI-BLAST. One of the reasons is Bayesian search predicted several false related pairs with very low Bayesian evidence. For example, at lowest Bayesian evidence value 0, there are 17 false positive pairs, equivalent to 0.0128 EPQ. In those cases, all the alignments generated by Bayesian search include a large aligned region which has an average length of 65 residues. The average rmsd of the 17 aligned regions is 11.5 Å. 3 out of 17 pairs are clearly wrong, the rest of pairs of structures have very similar secondary structures, but fold differently. For example, 1sria (2.40.1.1.1) and 1arb (2.29.1.1.1) both have four-strand beta-sheets, the four beta-sheets are packed differently in the two structures, shown in Fig. 3. The rmsd of the superposed structures is 9.9 Å. We cross-checked the 17 false positive pairs using VAST (Madej, Gibrat, & Bryant 1995; Gibrat, Madej, & Bryant 1996) . At least one pair, Succinyl-CoA synthetase 1scud2 (3.18.1.8.1) and D-lactate dehydrogenase 2dldal (3.13.9.1.4) are structural neighbors according to VAST.

Discussion

We compared Bayesian search with the Smith-Waterman method using the best parameter settings for structure comparison from Brenner *et al.* (1998), and PSI-Blast. From the above comparisons, Bayesian search identified more structure neighbors for EPQ >0.15.

Bayesian search and PSI-Blast use different sets of related sequences for model construction. The difference in finding remotely related sequences is not due to more distantly related sequences included in Bayesian search model, but due to difference in model construction method. Transitive Blast is similar to IS method except Fasta method is used in IS method. Park *et al.* (1998) shows that PSI-Blast has better sensitivity and selectivity than IS method. Thus, there could be more remotely related sequences used in PSI-Blast model than in Bayesian search model. More distant related sequences in a model may not guarantee a better model.

Park *et al.* (1998) shows that HMM performs better than PSI-BLAST. We excluded HMM in our compari-

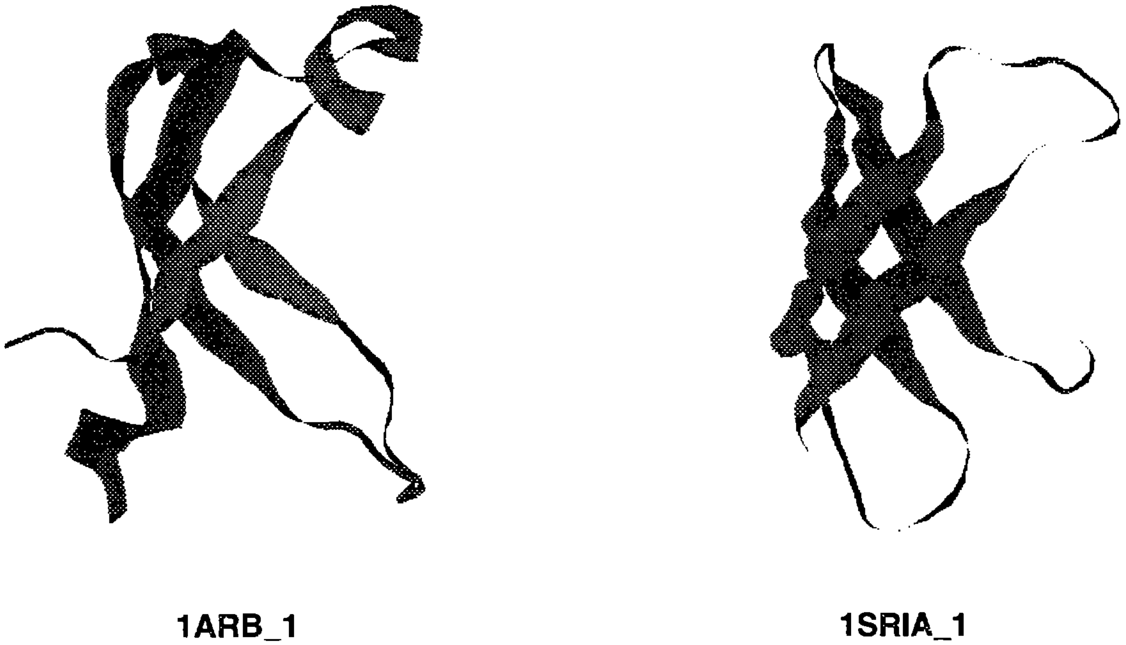

Figure 3: A false positive found by Bayesian search at low cutoff, Bayesian evidence equal to 0. The two structures, 1sria (residues 12-67) and 1arb (residues 21-76), have similar secondary structure elements, four beta-sheets. But the beta-sheets are folded in different ways.

son because it took too long to build HMM models for sequences in PDB40D-B. Normally it will take several hours to build a HMM model, while it only takes several minutes to build a Bayesian search or PSI-Blast model.

There are many sequence weighting methods (Lüthy, Xenarios, & Bucher 1994; Henikoff & Henikoff 1994) for constructing motif or profile models. In Bayesian search model construction, we assume that each sequence has identical weight because we purge closely related sequences. Neuwald *et al.* (1997) shows that when closely related sequences are removed, the identical weighting yields essentially equivalent results as using other weighting methods. To veryfying this assumption, we used ClusterW (Thompson, Higgins, & Gibson 1994a) to align the sequences used in our model construction. The sequence weights calculated using tree weights are identical for all sequences. We also applied a modified Henikoff and Henikoff (Henikoff & Henikoff 1994)'s weighting method to the alignments used in Bayesian search model construction. Xs (unaligned residues) in a column were treated as different from each other. The resulted sequence weights are flat. Both tests indicate that our assumption is reasonable. In addition, we tested $P(k > 0|Q, R)$, the posterior probability of two sequences having aligned blocks, as sequence weight function. The result is similar to the one using identical weighting.

Also during constructing models, aligned sequences can be so divergent that part of the alignment may not be reliable. For all the multiple sequence model methods we compared, all residues in an aligned column have the same weight. Some expertise may be needed to get rid of unreliable portions of alignments. In the Bayesian search model, each residue in an aligned column has different weight proportional to, $P(A_{i,j} = 1|Q, R)$, the marginal posterior alignment distribution which indicates how reliable the alignment of the pair is. Thus, the doubtful portions of alignments will be weighted down automatically. This is a distinguishing feature of the Bayesian search method. Also the whole column can be weighted down if $\sum_n P(A_{i,j} = 1|Q, R^{(n)})$, the probability that the site in the query sequence is conserved, is low.

In conclusion, we tested the Bayesian search along with pairwise and multiple sequence comparison methods on structural databases. Bayesian models can connect more remotely related proteins than other methods.

The software is available at request to authors. Also they can be download from http://www.wadsworth.org/res&res/bioinfo/.

Aknowledgements

We thank Jun S. Liu for helpful discussion. We also thank Computational Biology Group at Amgen, Inc. for providing all the computational resources.

References

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; and Lipman, D. J. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acid Research* 25(17):3389–3402.

Baldi, P.; Chauvin, Y.; Hunkapiller, T.; and McClure, M. A. 1994. Hidden markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA* 91:1059–1063.

Boriack-Sjodin, P.; Margarit, S.; Bar-Sagi, D.; and Kuriyan, J. 1998. The structural basis of the activation of ras by sos. *Nature.*

Brenner, S. E.; Chothia, C.; and Hubbard, T. J. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA* 95:6073–6078.

Day, G.; Mosteller, R.; and Broek, D. 1998. Distinct subclasses of small gtpases interact with guanine nucleotide exchange factors in a similar manner. *Mol Cell Biol.*

Eddy, S.; Mitchison, G.; and Durbin, R. 1995. Maximum discrimination hidden markov models of sequence consensus. *J Comput Biol* 2(1):9–23.

Eddy, S. 1995. Multiple alignment using hidden markov models. *Ismb* 3:114–120.

Gibrat, J.; Madej, T.; and Bryant, S. 1996. Surprising similarities in structure comparison. *Curr Opin Struct Biol* 6:377–385.

Gribskov, M.; McLachlan, A. D.; and Eisenberg, D. 1987. profile analysis: Detection of distantly related proteins. *Proc. Natl. Sci. USA* 84:4355–4358.

Henikoff, S., and Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Matl. Acad. Sci. USA* 89:10915–10919.

Henikoff, S., and Henikoff, J. G. 1994. Position-base sequence weights. *J. Mol. Biol.* 243:574–578.

Kawashima, T.; Berthet-Colominas, C.; M, M. W.; Cusack, S.; and Leberman, R. 1996. The structure of the escherichia coli ef-tu.ef-ts complex at 2.5 a resolution. *Nature* 379:511–518.

Krogh, A.; Brown, M.; Main, I. S.; Sjölander, K.; and Haussler, D. 1994. Hidden markov models in computational biology: Applications to protein modeling. *J Mol Biol* 235:1501–1531.

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.

Lüthy, R.; Xenarios, I.; and Bucher, P. 1994. Improving the sensitivity of the sequence profile method. *Protein Science* 3:139–146.

Madej, T.; Gibrat, J.; and Bryant, S. 1995. Threading a database of protein cores. *Proteins* 23:356–69.

Neuwald, A. F.; Liu, J. S.; Lipman, D. J.; and Lawrence, C. E. 1997. Extracting protein alignment model from the sequence database. *Nucleic Acids Research* 25(9):1665–1677.

Pai, E.; Kabsch, W.; Krengel, U.; Holmes, K.; John; and Wittinghofer, A. 1990. Structure of the guanine-nucleotide-binding domain of the ha-ras oncogene product p21 in the triphosphate conformation. *Nature* 341:209–214.

Park, J.; Teichmann, S. A.; Hubbard, T.; and Chothia, C. 1997. Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.* 273:349-354.

Park, J.; Karplus, K.; Barrett, C.; Hughey, R.; Haussler, D.; Hubbard, T.; and Chothia, C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* 284:1201–1210.

Pearson, W. R. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85:2444–2448.

Sankoff, D. 1972. Matching sequences under deletion/insertion constraints. *Proc. Acad. Sci. USA* 69:4–6.

Schwartz, R., and Dayhoff, M. 1978. *Atlas of Protein Sequence and Structure*, volume 5. National Biomedical Research Foundation, Washington, DC. chapter Matrices for detecting distant relationships, 353–358.

Smith, T. F., and Waterman, M. S. 1981. Indentication of common molecular subsequences. *J. Mol. Biol.* 147:195–197.

Thompson, J.; Higgins, D.; and Gibson, T. 1994a. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22:4673–4680.

Thompson, J.; Higgins, D.; and Gibson, T. 1994b. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput Appl Biosci* 10(1):19–29.

Zhu, J.; Liu, J. S.; and Lawrence, C. E. 1998. Bayesian adaptive sequence alignment algorithms. *Bioinformatics* 14(1):25–39.