ORIGINAL RESEARCH

Appl Bioinformatics 2004; 3 (2-3): 137-148
1175-5636/04/0002-0137/$31.00/0

A Sequence Alignment-Independent Method for Protein Classification

John K. Vries,[1] *Rajan Munshi*,[1] *Dror Tobi*,[1] *Judith Klein-Seetharaman*,[2] *Panayiotis V. Benos*[3] and *Ivet Bahar*[1]

1 Department of Molecular Genetics and Biochemistry, School of Medicine, Center for Computational Biology and Bioinformatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
2 Department of Pharmacology, School of Medicine, University of Pittsburgh, and Language Technologies Institute, Institute for Software Research International School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
3 School of Medicine, Department of Human Genetics, Graduate School of Public Health, Center for Computational Biology and Bioinformatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Abstract

Annotation of the rapidly accumulating body of sequence data relies heavily on the detection of remote homologues and functional motifs in protein families. The most popular methods rely on sequence alignment. These include programs that use a scoring matrix to compare the probability of a potential alignment with random chance and programs that use curated multiple alignments to train profile hidden Markov models (HMMs). Related approaches depend on bootstrapping multiple alignments from a single sequence. However, alignment-based programs have limitations. They make the assumption that contiguity is conserved between homologous segments, which may not be true in genetic recombination or horizontal transfer. Alignments also become ambiguous when sequence similarity drops below 40%. This has kindled interest in classification methods that do not rely on alignment. An approach to classification without alignment based on the distribution of contiguous sequences of four amino acids (4-grams) was developed. Interest in 4-grams stemmed from the observation that almost all theoretically possible 4-grams (20^4) occur in natural sequences and the majority of 4-grams are uniformly distributed. This implies that the probability of finding identical 4-grams by random chance in unrelated sequences is low. A Bayesian probabilistic model was developed to test this hypothesis. For each protein family in Pfam-A and PIR-PSD, a feature vector called a probe was constructed from the set of 4-grams that best characterised the family. In rigorous jackknife tests, unknown sequences from Pfam-A and PIR-PSD were compared with the probes for each family. A classification result was deemed a true positive if the probe match with the highest probability was in first place in a rank-ordered list. This was achieved in 70% of cases. Analysis of false positives suggested that the precision might approach 85% if selected families were clustered into subsets. Case studies indicated that the 4-grams in common between an unknown and the best matching probe correlated with functional motifs from PRINTS. The results showed that remote homologues and functional motifs could be identified from an analysis of 4-gram patterns.

The amount of genetic information deposited in public protein databases such as Swiss-Prot/TrEMBL,[1] PIR-PSD,[2] RefSeq,[3] GenPept[4] and the Protein Data Bank (PDB)[5] has increased exponentially with the advent of the genome era.[6-13] The PIR-NREF database,[2] combining nonredundant sequences from all of these sources, currently contains 1 292 569 entries. Understanding the structure and function of these newly discovered sequences is the key to understanding and treating disease.[14,15] The rate of accumulation of these new sequences, however, is far beyond the capacity of the scientific community to determine their attributes through biochemical or crystallographic methods. Many databases of protein families have been developed in the public domain to

facilitate classification by similarity. These include Pfam,[16] PIR-PSD,[2] PROSITE,[17] BLOCKS,[18] PRINTS,[19] SMART,[20] ProDom[21] and CDD.[22] Links between these databases are provided by InterPro.[23]

The most popular approaches for determining homology are based on sequence alignment.[24,25] Possible alignments between two sequences are scored using substitution matrices[26] derived from evolutionary studies such as PAM[27] or multiple sequence alignments such as BLOSUM.[28] Scores for each potential alignment are treated as log-likelihood ratios and added. Scores are normalised using statistical approaches related to the extreme value distribution.[25,29] Sequences with optimal alignments below a particular E-value (typically 0.001) are considered significant.[25,29] Optimal global and local pathways are determined using dynamic programming techniques.[30-32] Common algorithms for pairwise comparisons of protein sequences include FASTA,[7,9,33] which takes advantage of high-scoring *n*-grams (usually called *k*-tuples), and BLAST®,[24] which extends high-scoring triplets. PSI-BLAST,[25] an enhancement of BLAST® that utilises position-specific scores calculated from highest scoring BLAST® hits on successive iterations, is particularly useful for finding distantly related proteins.

Most of the large protein family databases in the public domain are constructed from multiple alignments using profile hidden Markov models (HMMs).[34-39] The largest of these is Pfam, which currently covers 93% of the sequences in the Swiss-Prot/TrEMBL[1] database. A subset of Pfam called Pfam-A is based on seed alignments that have been verified by human experts.[40] It currently contains 975 024 domain sequences organised into 6193 families. Attempts have also been made to align sequences without training sets in order to build evolutionary trees directly. The concept is to build a guide tree using pairwise comparison and to refine it with sequence-family and family-family comparisons. The most prominent example is ClustalW.[41]

Alignment-based methods have limitations, however.[42] They are based on the assumption that contiguity is conserved between homologous segments,[42] which may not be true in genetic recombination or horizontal transfer.[43,44] Alignments also become ambiguous when sequence similarity drops below 40% and become unusable when this level reaches 20–25%.[45-47] This has led a number of investigators to explore classification programs that do not rely on sequence alignment.[42]

The majority of alignment-independent approaches are based on the statistical properties of *n*-grams. The vector space approach treats each protein sequence as a feature vector of *n*-grams. Similarity between sequences is determined using a variety of metrics including Euclidean distance,[48-51] covariance[52-54] and the cosine between feature vectors.[55] Covariance-based metrics such as the Mahalanobis distance have also been employed.[56] They add computational complexity to the approach but they make comparisons independent of scale.[42] Other approaches exploiting the statistical properties of *n*-grams involve information theory and the Kullback-Liebler discrepancy or relative entropy.[57,58] Approaches not involving *n*-grams have also been explored. These include universal sequence maps[59,60] based on chaos theory, compression methods utilising Kolmogorov complexity[61] and a variety of machine learning algorithms including support vector machines[62-67] and neural nets.[68-71] Alignment-independent methods have been used to advantage to pre-process large database searches, but they have not achieved wide popularity as primary tools despite their potential utility.[42]

We developed an approach to classification without alignment (CWA) based on the distribution of 4-grams in protein sequences. Interest in 4-grams stemmed from the observation that almost all theoretically possible 4-grams (20^4) occur in natural sequences and the majority of 4-grams are uniformly distributed. This implies that the probability of finding the same 4-grams by random chance in two unrelated sequences is low. A Bayesian probabilistic model was developed to test this hypothesis. The model was used as the basis of a classification system for detecting protein family homologues. The potential to identify functional motifs through analysis of 4-gram patterns was explored in a series of case studies.

Methods and Results

Database Selection and the Distribution of 4-Grams

Two databases were selected for analysis. The first database was Pfam-A[16] (release 7.7), which is a database of protein families at the domain level. The domains were derived from Swiss-Prot[1] (release 41.0) containing 122 564 sequences and TrEMBL[1] (release 23.0) containing 830 524 sequences. The domains were identified using HMMs trained from curated seed alignments.[29,34,72-75] This release contained 975 024 domains organised into 5193 families for an average of 188 members per family. The second database was PIR-PSD (release 76.00), which is a database of 283 308 nonredundant whole protein sequences. A 158 938-member subset of these sequences was organised into 19 559 PIR superfamilies[45] for an average of eight members per superfamily.

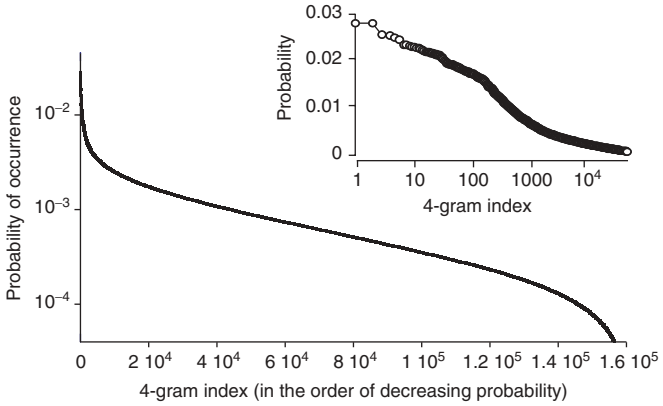

Fig. 1. Probability of occurrence of all the 4-grams in Pfam-A. A total of 159 996 unique 4-grams were found in 975 024 domain sequences. The 4-grams are displayed in rank order by decreasing probability, p*iU*. p*iU* is below 0.25% for a large majority of 4-grams. The inset displays the subset of the most probable 4-grams. The most common 4-grams occurred in less than 3% of sequences. This suggests that the probability of finding a large number of matching 4-grams in unrelated sequences is low.

The probability of occurrence of different 4-grams was determined by enumerating all possible 4-grams in Pfam-A and PIR-PSD. The overall 4-gram count as well as the count of unique 4-grams per sequence was recorded. The distribution of unique 4-grams per sequence for Pfam-A is shown in figure 1. The x axis refers to the 4-gram serial numbers, in the order of descending probability of occurrence in the database. Any 4-gram containing the unknown residue type X was discarded. The ambiguous (and rare) residue types B (Asx) and Z (Glx) were mapped to D and E, respectively. Theoretically, there are 160 000 (20^4) possible 4-grams. Pfam-A contained 159 996. The most common 4-grams were GLLL and NNTR with probabilities of 2.8% (see figure 1). These were succeeded by GDIR and CTRP (2.6% and 2.5%). In the remainder of this article this background distribution will be referred to by the letter *U* (universal set). The results for PIR-PSD were comparable.

Figure 1 implies that the probability of finding multiple 4-grams in common between unrelated sequences is low. To confirm this, pairs of sequences were randomly selected with a Monte Carlo (MC) algorithm from amongst the 975 024 sequences in Pfam-A, and the number of matching *n*-grams was recorded for each pair. The probability of pairs having *m* identical 4-grams (or the random/background probability of finding *m* identical 4-grams) was determined for $m = \{1 \ldots 8\}$. The results are shown in table I. These values define the *a priori* probabilities of occurrence of *m* matches p*m*(*x*,*y*|*U*) in the examined database (Pfam-A). We note that the percentage probability of randomly finding five matching 4-grams, for example, is 0.146%. These percentages would be an order of magnitude smaller if matches from members of the same family were excluded from the simulation.

The ability to use 4-grams to identify members of protein families is a function of the difference in 4-gram distributions between family members and nonfamily members. The Kullbak-Liebler distance or relative entropy provides a convenient method for quantifying this difference.[57,58] If *P* and *Q* are two different distributions, the relative entropy is defined as (equation 1):

$$H(P||Q) = \sum \mathrm{p}_P^i \log \frac{\mathrm{p}_P^i}{\mathrm{p}_Q^i}$$

where $\mathrm{p}^i P$ and $\mathrm{p}^i P$ are the probabilities of the *i*th 4-gram in the respective distributions *P* and *Q*. If the distribution of *Q* is uniform, the relative entropy is equivalent to the difference in information content between the two distributions.

The relative entropy of each of the 5193 domain families defined in Pfam-A was determined with respect to *U*. The results are shown in figure 2. The figure displays the probability distribution (or the number of occurrences) of relative entropies. Grids of size 0.8 are considered along the x axis. The relative entropies were normalised for sequence length. The x axis values thus reflect the information gain per 4-gram; there is an information gain for all families. The relative entropy is >1 per 4-gram in two-thirds of the examined domains, indicating about 1 order of magnitude difference in population of the probe 4-grams in the families compared with the background. These results suggest that differences in 4-gram distributions could be used for identifying family membership.

Theoretical Basis of the Bayesian Probabilistic Model

A Bayesian probabilistic model for comparing sequences based on the distribution of 4-grams was developed by adapting the mathematical framework developed by Durbin et al.[29] for sequence alignment. Let *x* and *y* be two sequences that have *m* *n*-grams in common. The probability of *m* *n*-grams in common given

Table I. Probability of identical 4-grams in unrelated sequences

Matches (*m*)	1	2	3	4	5	6	7	8
Probability	0.12719	0.03459	0.01092	0.00380	0.00146	0.00062	0.00026	0.00012

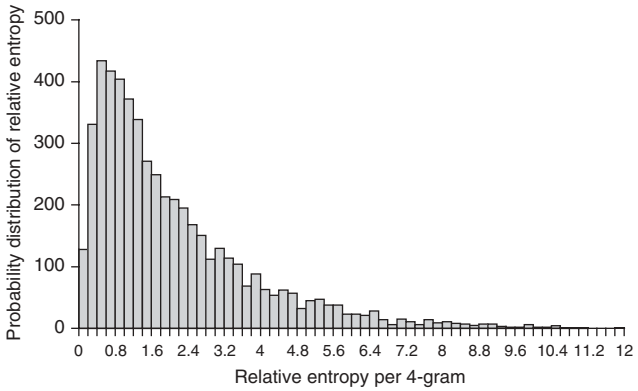

Fig. 2. Probability distribution of the relative entropy per 4-gram for the 5193 families in Pfam-A. There is a significant gain in information for the majority of family distributions compared with the background distribution. The mean value is calculated to be 2.072, indicating that the preferences for the 4-grams in the Pfam families are about 2 orders of magnitude different from those in the background distribution. Note that the information gain per sequence is the x axis multiplied by the sequence length, because the results are normalised with respect to sequence lengths (i.e. divided by the total number of *n*-grams in each sequence).

that the sequences belong to the same homologous family F is given by (equation 2):

$$p_m(x,y|F) = \prod p^i_F$$

The product Π is performed for $1 \leq i \leq n$. The same probability by random chance (or in U) is (equation 3):

$$p_m(x,y|U) = \prod p^i_U$$

such that the odds ratio of the two likelihoods can be expressed as (equation 4):

$$p_m(x,y|F)/p_m(x,y|U) = \prod(p^i_F/p^i_U)$$

Taking the log of both sides yields an additive scoring function similar to the one described by Durbin et al.[29] Let the log-odds score be defined as (equation 5):

$$\frac{p_m(x,y|F)p(F)}{[p_m(x,y|F)p(F) + p_m(x,y|U')p(U')]}$$

which simplifies to equation 6 after division by $pm(x,y|U')p(U')$ and substitution of the scoring function.

$$p_m(x,y|F) = \frac{e^{S'}}{(1 + e^{S'})}$$

where (equation 7):

$$S' = S_m(x,y) + \log \frac{p(F)}{p(U)}$$

The *a posteriori* probability calculated with equations 6 and 7 can be used as a similarity metric between two sequences based on common 4-grams. It can also be used to determine the similarity between an unknown sequence and the 4-gram distribution characterising a protein family.

Probe Creation and Certification

Feature vectors called probes were created for 5174 of the 5193 domain families in the 7.7 release of Pfam-A. Nineteen sequences were excluded because they had less than ten unique 4-grams or <20 members in their families. Probe construction involved three steps. The first step was to calculate the log-likelihood ratio of each 4-gram in the family and the corresponding 4-gram from the background distribution. The second step was to sort this list in descending order. The third step was the determination of the probe vector size. The concept was to limit the probe to the most discriminative 4-grams to reduce the dimensionality of the probe while retaining the ability to identify family members. To this aim, we increased the probe size in increments of ten 4-grams starting at the top of the rank-ordered list. After each increment, the probability of finding three matches between the probe and an equal size 'random' probe constructed from the background distribution minus the family distribution was determined by MC simulation. The process terminated when the probability of three matches by random chance reached 0.001. For Pfam-A, the number of 4-grams in a probe ranged from 10 to 480. The average probe contained 126 4-grams. Note that each score contributing to the Bayesian expression is a log-likelihood ratio, and thus the *n*-gram probe vectors are weighted such that matches to the most commonly occurring of the dominant 4-grams were scored as more significant.

After all probes were created, an MC simulation was run to determine the probability of m hits between each probe and its family distribution. This was repeated for the background distribution minus the family distribution. The same Bayesian probabilistic model used to measure similarity was applied to the log odds ratio of the probability of m hits given the family, $p(m|F)$, versus the probability of m hits given nonfamily, $p(m|U)$, to yield the p-value as a function of m. The results for all probes were formatted into a look-up table to provide p-values based on the number of common 4-grams between a probe and an unknown. The MC certification process is depicted in figure 3. The table in this figure shows the average probability for $m = \{1\ldots6\}$ over all 5174 Pfam-A families. The p-value reaches 0.05 on average (0.95–1.0) when the common number of 4-grams reaches five.

For all the families in Pfam-A, $p(F|m)$ was calculated for $m = \{1\ldots5\}$. The results are presented in figure 4. The most important feature in this figure is again the size of m when the probability of

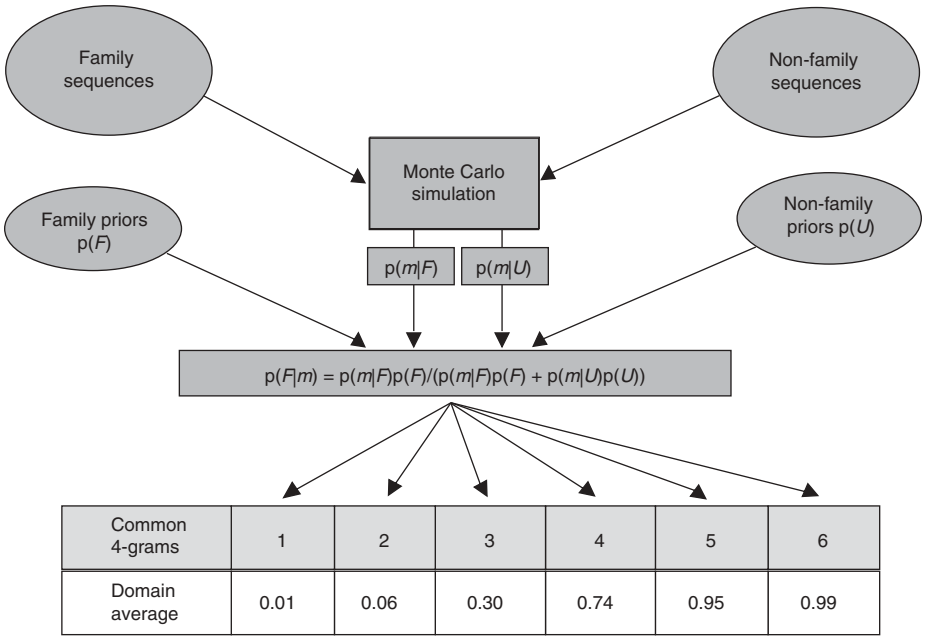

Fig. 3. Schematic description of the Monte Carlo calculations for determining the conditional probability, p(*F*|*m*), of family *F*, also called probe certification value, given *m* common *n*-grams between the probe and query sequence.

a correct classification approaches 1.0. A small percentage achieved this with $m = 2$ or $m = 3$. Almost 90–95% reached this by the time $m = 5$. A small minority never approached 1.0 even with larger *m*. These represent family distributions that do not differ significantly from the background distribution. The average domain size was calculated to be 145 residues. The case of five matching 4-grams thus corresponds to only 6–14% of the average sequence, depending on the extent of overlap between *n*-grams. This number (10 ± 4) represents a lower bound for the sequence identity between the pairs having $m = 5$ matching *n*-grams, bearing in mind the two sequences may also have shorter *n*-grams in common.

Jackknife Testing for Pfam-A

The specificity and sensitivity of CWA for protein domains were estimated by dividing Pfam-A version 7.7 into equal training and testing sets. Version 7.7 contains 975 024 sequences in 5193 domain families. Sequences containing less than ten 4-grams were rejected, as were families having <20 members. The requirement for ten 4-grams eliminated most of the short repeat domains. The final set contained 949 835 sequences and 3671 families. These were assigned to training and test sets on a family-by-family basis under control of a random number generator. The training set contained 474 918 members and the test set 474 917 members.

The training set was used to create 3671 probe models. Each test (query) sequence was converted to a probe and compared with each of the 3671 probe models. The probability that the test sequence was a member of each family was calculated using equation 6 and equation 7, in which *x* refers to the query sequence and *y* to the family probe vector. The final list of 3671 probabilities calculated for the query sequence was sorted into descending order. If the correct Pfam family was associated with the sequence in first place, the result was labelled a true positive (TP). If the sequence in first place was not the correct sequence, the result was labelled a false positive (FP). Calculations were repeated for all the 474 917 sequences in the testing set. Confidence limits were

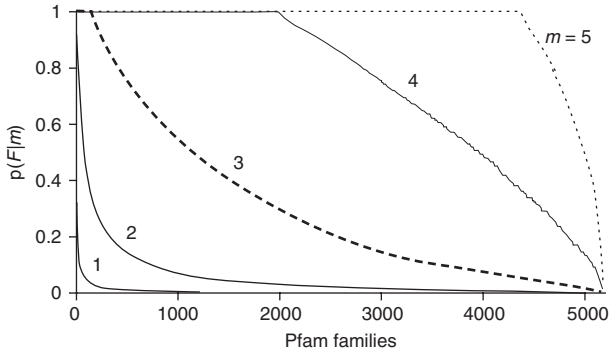

Fig. 4. The posterior probability of a correct classification into a given Pfam-A family *F*, given that the query sequence has *m* common 4-grams with the probe vector that represents the particular family. The x axis refers to the 5174 Pfam-A domain families rank ordered in decreasing probabilities. Results are shown for five cases, $m = 1 \ldots 5$. Correct classification reaches the 90% level when the number of 4-grams in common reaches five.

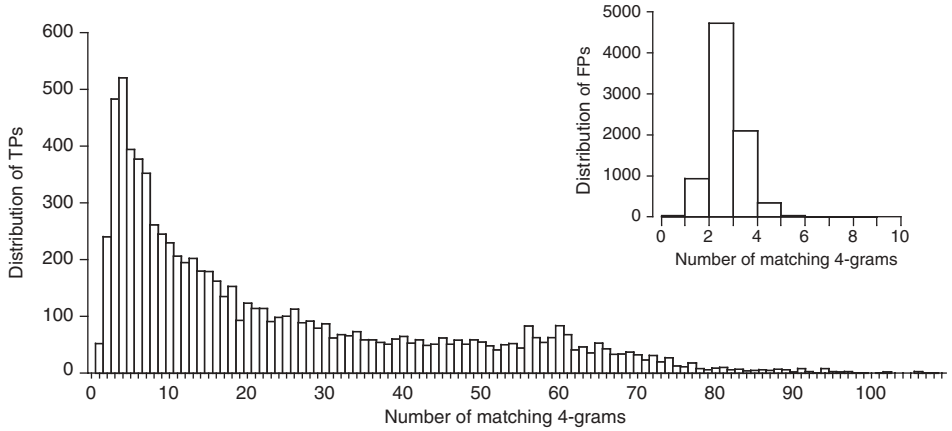

Fig. 5. Distribution of true positives (TPs) as a function of the number of matching 4-grams between the 4-gram vectors of the query sequence and the family probe vectors. The inset shows the distribution of false positives (FPs) as a function of the number of matching 4-grams. The distributions are generated by randomly sampling the 327 741 TPs and 147 176 FPs obtained from the application of a 4-grams-based family classification jackknife test to half of the Pfam-A sequences, the other half being used for defining/training the probe vectors of the Pfam-A families.

assigned using the Bayesian probabilistic model depicted in figure 3. A look-up table relating the probability of being correct as a function of the number of 4-grams in common was constructed for each of the 3671 probe models. The results showed that 69% of sequences were correctly classified. Correctly classified sequences (TPs) showed a median of 15 4-gram matches with their family probe, which corresponds to 12–41% sequence identity for an average Pfam-A domain of 145 residues, depending on whether the 4-grams shared common residues. FPs showed an average of only 2.35 4-gram matches. The distributions of the matching 4-gram counts (m) significantly differ in the TPs and FPs as shown in figure 5. The average confidence limit for the TPs was 0.047. The average confidence limit for the leading contender for the FPs by contrast was 0.57. Analysis of FPs with poor confidence limits showed disproportionate membership in large Pfam families. For large families, probe sizes, which averaged 126 4-grams in this study, are insufficient to characterise the family 4-gram distribution. This implies that large families need to be separated into subfamilies. The analysis of FPs in the section Analysis of False Positives supports this hypothesis. Selection of FPs with good confidence limits (p-value < 0.05) yielded 10 667 instances from 147 176 FPs (8%). Analysis of this group showed mainly proteins with multiple domains where one domain was favoured over the other or classification near the top of the list, but not in the first position. PIR-PSD results (not included in this article) were comparable (71% success rate).

Analysis of False-Positives

Sixty-nine percent of the sequences in Pfam-A and 71% in PIR-PSD were correctly classified using the 4-gram-based CWA. This implies an ~30% error rate in Pfam-A and PIR-PSD. Four hypotheses were postulated to explain this error rate: (i) there may be unrecognised *subclasses* within families that could be detected by probes with greater specificity; (ii) evolutionary relationships might be so remote that sequence identities have fallen below the level detectable by n-grams; (iii) the conserved portions of some sequences may contain regions that are below the resolution of a 4-gram (an example might be zinc finger proteins, which are characterised by a motif containing two noncontiguous histidines, two noncontiguous cysteines and a set of wildcards[17]); and (iv) the Pfam-A classification may contain erroneous alignments. The seed alignments for the profile HMMs used in Pfam are carefully curated, but the full alignments are produced automatically.[16] Some insights into this process were gained by examining the relative entropies associated with the TPs and FPs founds in the jackknife tests for Pfam-A. Figure 6 shows the results for all the

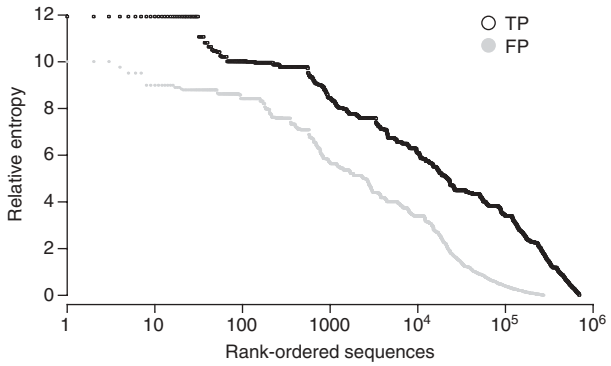

Fig. 6. Comparison of the relative entropy per 4-gram of the true positives (TPs) and false positives (FPs). FPs with high values suggest distributions that might be exploited by breaking families into subclasses and creating probes with greater specificity.

Table II. Comparison of PSI-BLAST and classification without alignment (CWA) methods

Program	PSI-BLAST	CWA	Agreement
Successes (total)	57 (143)	54 (143)	48 (57)
Success rate[a]	0.399	0.378	0.842

a Where 1 is 100%.

Pfam-A sequences. Interestingly, some TPs show low relative entropies, while the entropies of some FPs are high. The former group may represent distant evolutionary relatives with 4-gram distributions close to the background noise but still detected by a few discriminative 4-grams. The FPs with high relative entropy, on the other hand, probably represent distributions that could be exploited by breaking members into subclasses and building probes with greater specificity. To evaluate this hypothesis, an internal consistency check was conducted on Pfam-A. A new set of family probes was constructed from the 272 633 FPs observed in the jackknife tests. A total of 38% were correctly classified as TPs in this second generation classification. This provides evidence that subclasses exist within certain families. Identification of these subclasses, or construction of probes that take account of these subclasses, would lead to probes with higher specificity and should boost the accuracy rate of the overall classification to the 80–85% accuracy-level range.

Comparison of Clarification Without Alignment with PSI-BLAST Similarity Searches

The jackknife tests with Pfam-A and PIR-PSD demonstrate that CWA can successfully classify protein sequences. Models based on Pfam-A or PIR-PSD, however, inherit any potential flaws in the parent models defining the family members. To avoid this, a new series of studies is underway using release 1.29 of PIR-NREF containing 1 292 569 sequences.[2] In these studies, each protein was declared to be its own model. The feature space was also extended to include all possible 4-grams in a window of six that allowed one or two gaps. Using an inverted index with *n*-grams as keys and NREF identification numbers as values, a list of classification candidates rank ordered by *n*-grams in common can be built for any unknown sequence. An E-value can be assigned to each candidate on the list based on the probability of *m n*-grams in common by random chance[29] using the same theoretical foundation as the Pfam-A and PIR-PSD studies. The efficacy of this methodology was tested by classifying the 143 open reading frames (ORFs) in the virulence plasmid of anthrax reported by Okinaka et al.[76] The threshold for successful classification was a protein with annotation confirming function and an E-value better than −10. Initially, the results were compared with the results reported by Okinaka et al.[76] in the literature. We repeated the study of Okinaka et al.[76] using PSI-BLAST for all 143 ORFs using the same criteria as CWA. The results of the studies using identical criteria are presented in table II.

Table II shows that PSI-BLAST and CWA successfully classify 40% of unknown protein sequences in terms of assignment of function. The classifications agree with each other in 84% of cases. If the criteria are relaxed to allow recognition of conserved domains of unknown function, the success rate increases to 72%. This methodology is still being refined, but the initial results indicate it produces results comparable to PSI-BLAST under stringent conditions.

Case Studies: Biological Meaning of the Most Discriminative 4-Grams

Case studies relating the most discriminative *n*-grams to known biological and structural features were conducted for selected members of the serine protease, protein kinase and G-coupled protein receptor (GPCR) families. A detailed analysis of the GPCR serotonin receptor 5h1a_human (P08908) follows. Similar results were obtained for the selected serine proteases and the protein kinases. The receptor 5h1a_human was correctly classified into its Pfam-A family with a 4-gram probe vector of $d = 22$ elements in common. These 22 *n*-grams, referred to as the most discriminative *n*-grams, are illustrated in figure 7 together with the conserved motifs from the PRINTS database.[77] Before interpreting these results, we briefly review the known structural and functional features of the GPCR family, and in particular the class A GPCRs to which the examined protein belongs. Rhodopsin, the mammalian dim light photoreceptor molecule, is the defining member of the largest subfamily of GPCR, the class A rhodopsin-like receptors.[78] GPCRs share a common structural motif, a bundle of seven transmembrane helices (see figure 7), which divides the proteins into extracellular (EC), transmembrane (TM) and cytoplasmic (CP) domains. The first step in GPCR signal transduction is the binding of ligands (agonists) specific to each receptor. Ligand-binding occurs in the EC and/or TM domains, and both domains are structurally tightly coupled. This coupling

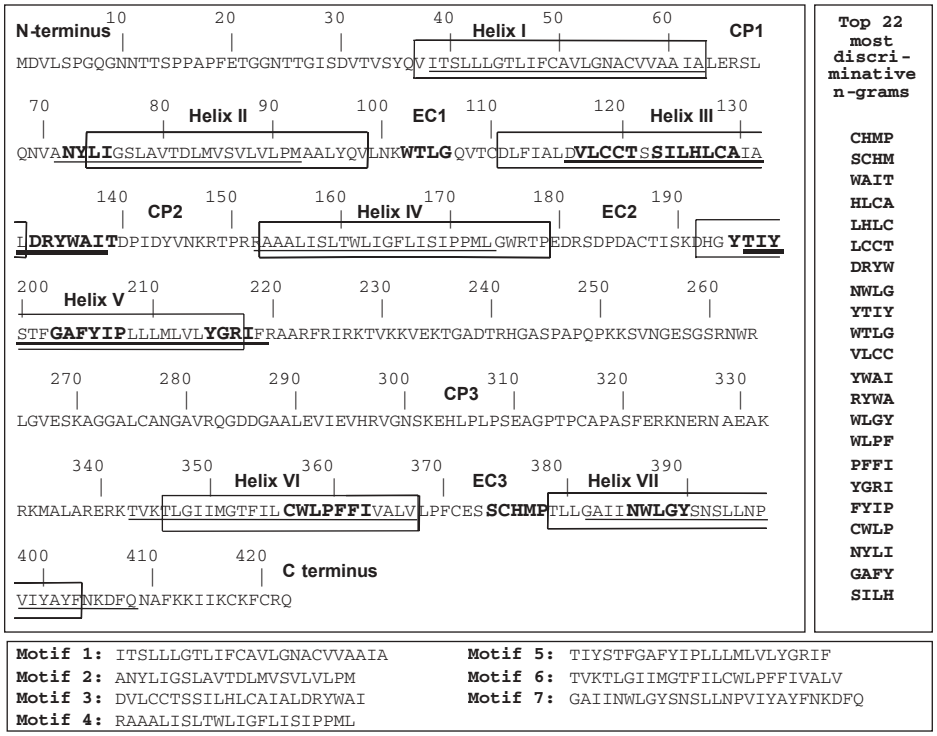

Fig. 7. Distribution of 4-grams (bold letters) in the human serotonin receptor (ID: 5h1a_human) in comparison with the position of PRINTS motifs (underlined) and transmembrane helices (boxes). Extracellular loops are indicated by the N-terminus and EC1 to EC3, cytoplasmic loops are indicated by the C-terminus and CP1 to CP3. The seven transmembrane helices are indicated as helix I to helix VII. The 4-grams shown are the top 22 most strongly contributing to the classification. They are also listed on the right of figure, rank ordered by weight. The motifs from the PRINTS database (http://bioinf.man.ac.uk/dbbrowser/PRINTS) are listed at the bottom of the figure.

has been studied extensively in rhodopsin, where it was shown that point mutations in the TM or EC domains known to induce retinal degeneration cause misfolding via disruption of a disulphide bond between a cysteine at the end of helix III and a cysteine in the second EC loop.[79,80] All EC loops are intimately involved in providing the environment for the disulphide bond and the interface to the centre of the TM bundle, as confirmed in the crystal structure of rhodopsin.[81]

Ligand-binding, and in the special case of rhodopsin, retinal isomerisation, induces an opening of the helical bundle towards the CP ends of the helices by an outward movement of helices II, III, VI and VII, which behave as rigid bodies.[82-93] This movement results in increases in distances between the CP loops connecting the helices, most notably CP1 and CP4[85,86] and CP2 and CP3[90] upon light-activation. The resulting cleft in the centre of the helical bundle is believed to expose a highly conserved DRY (or ERY) sequence in CP3 that has been shown to be critical for the interaction with the G protein.[94] Preventing this exposure by cross-linking inhibits activation of the G protein.[94]

A number of the above structural features are conserved across the GPCR family. For instance, within the class A subfamily: (a) the seven helical motif defines the family, and the highest conservation of amino acid residues is found in this domain;[95] (b) the disulphide bond in the EC domain is conserved in ~95% of class A type GPCR; and (c) CP3, in which the highly conserved (D/E)RY sequence is located, is the dominant interaction site for binding of the G protein. The conservation of these features supports the hypothesis that similar helix movements occur in all GPCRs in response to ligand-induced activation.[95] Therefore, the mechanism of GPCR activation is believed to be fundamentally the same for all GPCRs.

We asked if these general features that have been identified by extensive studies of the GPCR family, and in particular rhodopsin, are consistent with the top 22 most discriminative 4-grams of our classification system. These 4-grams are shown in bold font in figure 7 together with the set of the motifs (or fingerprints) from the PRINTS database, as well as the seven TM helices. PRINTS located motifs at the locations of the seven TM helices, since these

are the most highly conserved regions in GPCR. The discriminative *n*-grams identify these to some degree also (in fact, the majority of 4-grams is located within PRINTS motifs), but interestingly, most of the 4-grams are within helix III, the helix that is most buried in the seven helical cluster.[81] Furthermore, the 4-grams were more specific in the identification of functional sites than PRINTS. There are no *n*-grams located in helices I and V, two helices that do not participate in the cooperative ligand-induced opening of the TM domain. While PRINTS does identify the (D/E)RY motif, it does so only within a longer stretch of amino acids that includes the entire helix III with additional functions. The *n*-grams, in contrast, identified the (D/E)RY region as a distinctive location, suggesting separate functions from 4-grams within helix III. Furthermore, 4-grams identified additional features important for GPCR that PRINTS was not able to detect. These are the *n*-grams WTLG in EC1 and SCHMP in EC3. As described above, the EC loops are extremely important for the structural coupling between EC and TM domains. This coupling is an integral part of ligand-binding and signal transduction.

These results clearly suggest that the information used by homology detection and by feature identification in our classification system is to some degree overlapping, but not entirely. Homology detects features in addition to 4-grams, while 4-grams detect features in addition to homology. The detailed comparison described above also indicates that we will be able to improve detection of features by *n*-gram classification by including equivalent classes and similarity matrices of amino acids.

Discussion

Numerous classification schemes have been developed in the past based on the statistical properties of *n*-grams.[42,48-58] However, none have emphasised the properties of 4-grams. Our results showed that almost all 4-grams were represented in naturally occurring sequences and that the majority of 4-grams were uniformly distributed. Monte Carlo simulation demonstrated that the probability of finding 4-grams in common between unrelated sequences fell in the 0.05 range when the number of common 4-grams reached the 3–5 level. This number of common 4-grams corresponded to a sequence similarity in the 6–14% range, which was below the identity level discernible by BLAST®.[24,25] Examination of the relative entropy between the 4-gram distributions of the Pfam-A protein domain families and the PIR-PSD superfamilies compared with the background distribution showed a difference on average of two orders of magnitude. This implied that the statistical properties of 4-grams might be exploited for classification purposes. This hypothesis was tested using all the families in Pfam-A and PIR-PSD. Each family was divided randomly into equal test and training sets. Feature vectors of 4-grams called probes were created for all training sets, and probes were created for each sequence in the test set. The *a posteriori* probability that a test probe belonged to a family was calculated for each family in the training set. The final result was presented as a list of training models rank ordered by *a posteriori* probability. If the training model in first place on the list represented the correct classification it was declared a true positive. Correct results were obtained in 69% of cases for Pfam-A and 71% of cases for PIR-PSD. These results indicate that probes consisting of 4-grams have potential as classifiers under stringent conditions.

Only the 4-grams with the highest log-likelihood ratios with respect to the background distribution were placed in probes. The final size of a probe was truncated at the point where the chance for three 4-grams in common between unrelated sequences reached 0.001 as determined by Monte Carlo simulation. The average probe in Pfam-A had a size of 126 with a range of 10–480. Limiting the size of the average probe to 126 reduces the dimensionality of the computational task from 160 000 (20^4) to 126. This reduces the computational burden by more than three orders of magnitude. Calculating the *a posteriori* probabilities for a query sequence against 5174 Pfam-A family members on a mid-range Sun server, for example, required less than a second.

Recently Cheng et al. (unpublished observation) explored the use of unigrams, bigrams and trigrams in a naive Bayes classifier to classify GPCRs at the subfamily I and II levels. A chi-squared feature selector was used to extract the top 7400 *n*-grams from the $20^1 + 20^2 + 20^3 = 8420$ possibilities for subfamily level I. This gave an accuracy of 93.2%, outperforming by 4.8% the best result of Karchin et al.[66] using support vector machines. A similar approach yielded an accuracy of 92.4% for subfamily level II. Cheng et al. (unpublished observation) concluded that high accuracy could be obtained by selecting the most discriminative *n*-grams and that *n*-grams beyond trigrams were not necessary for robust classification. At first glance, these results seem at odds with our results. Closer examination reveals they are compatible. All 4-grams can be broken down into combinations of unigrams, bigrams and trigrams. The selection of the most discriminative *n*-grams in both approaches eliminates the *n*-grams with a high incidence and little discriminative power. Both reduce the dimensionality of the computational problem, although the 4-gram approach was one to two orders of magnitude more effective in this regard. Comparison of accuracy is not warranted here because the GPCR subfamily I and II levels represent a smaller and more

coherent dataset than all the sequences in Pfam-A and PIR-PSD. Unpublished classification studies by one of the authors (J. K. Vries) using 4-grams and sequences from the GPCRDB[96] database showed accuracy rates in the 90% range using the same jackknife protocol used for Pfam-A and PIR-PSD.

A problem with the classification results presented in this article is that they are dependent on the quality of the underlying models in the training sets. The family definitions in Pfam-A are derived from the application of an HMM to carefully curated seed alignments. However, the full family alignments are generated automatically. There is also a significant range in family size with some families having more than 16 000 members and others having only one or two. This problem was analysed by creating new training and test sets from the false positives from the classification runs for Pfam-A. New classification runs showed a successful classification rate of 38% for the original false positives. This implied that a significant number of families contained subsets of 4-grams related to family subdivisions. Normalising for the number of false positives in the primary runs, this implied that an accuracy rate in the 80–85% range might be achieved if the families could be grouped into their subfamilies.

In the classification studies presented in this article, the *a posteriori* probability was used as the metric. This was appealing because the theoretical foundation is similar to the foundation for alignment-based methods. This approach also enabled tables of p-values to be generated to judge the significance of individual classifications. The approach in this study is a variation of the vector space model. In theory, a wide variety of metrics could be applied. In the unpublished GPCR pilot study mentioned above (Vries JK, unpublished data), the cosine metric had 6.5% greater accuracy than the *a posteriori* probability. Other metrics will be explored in the future. The framework for this model is general. Any feature space and any mathematically sound metric could be used. Certification tables for the generation of p-values could also be created for any feature space and metric using the same methodology.

An advantage of 4-grams over combinations of unigrams, bigrams and trigrams is their ability to identify unique locally conserved regions in proteins. The redundancy of the smaller *n*-grams makes this difficult. If the 3% set of common 4-grams is excluded, the same 4-gram is rarely seen more than once in a sequence. The first case study presented in this protease case study demonstrates the correlation between discriminative 4-grams and two elements of the catalytic triad in the serine protease CTRA_BOVIN. It also demonstrates a significant correlation with the PRINTS motifs for this protein. Closer examination of the discriminative 4-grams for CTRA_BOVIN, however, shows that only two of the three members of the catalytic triad were identified. Unpublished studies by one of the authors (D. Tobi) using combinations of bigrams and trigrams with 4-grams showed the third member of the triad (Asp102) has a characteristic pattern that is below the resolution of a 4-gram but is picked up with trigrams. This suggests that combinations of bigrams, trigrams and 4-grams might have advantages. It also suggests that exploring the use of 4-grams in windows of 5–7 residues with gaps might be worthwhile since this would generate all possible subsets of unigrams, bigrams and trigrams as by-products. Similar results were obtained for the GPCR serotonin receptor 5h1a_human.

The results indicate that 4-grams have potential for the classification of unknown sequences. The most discriminative 4-grams seem to correlate with locally conserved functional motifs. The approach is also computationally efficient. It is general and can be modified to use a variety of feature spaces and metrics. The biggest limitation seems to be the quality of the underlying model defining the protein families.

Availability

The software developed for the studies in this article was written in Java. It is available as a jar file without charge to academic users. Interested parties should contact the primary author at vries@ccbb.pitt.edu for details.

Acknowledgements

The software developed for n-gram classification by JKV was supported by the National Institute of Standards and Technology (Advanced Technology Program grant number 70NANBOH3058). The work of PVB was supported by NSF grant number MCB0316255 and intramural funds of the Center for Computational Biology and Bioinformatics, the University of Pittsburgh Cancer Institute, and the Department of Human Genetics. This research was also supported by Information Technology grant number 0225636 from the National Science Foundation.

The authors have no conflicts of interest directly relevant to the content of this study.

References

1. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000; 28: 45-8
2. Wu CH, Yeh LS, Huang H, et al. The Protein Information Resource. Nucleic Acids Res 2003; 31: 345-7
3. Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 2001; 29: 137-40
4. GenPept Database. Genetic sequence data bank translated protein-coding sequences [online]. Available from URL: http://inn.weizmann.ac.il/databanks/genpept.html [Accessed 2004 Sep 21]
5. Bourne PE, Weissig H. Structural bioinformatics. Hoboken (NJ): John Wiley & Sons Inc, 2003: 181-198

6. Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. Nature 2002; 420: 520-62
7. Adams MD, Celniker SE, Holt RA, et al. The genome sequence of Drosophila melanogaster. Science 2000; 287: 2185-95
8. Gosele C, Hong L, Kreitler T, et al. High-throughput scanning of the rat genome using interspersed repetitive sequence-PCR markers. Genomics 2000; 69: 287-94
9. Holt RA, Subramanian GM, Halpern A, et al. The genome sequence of the malaria mosquito Anopheles gambiae. Science 2002; 298: 129-49
10. Kunst F, Ogasawara N, Moszer I, et al. The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature 1997; 390: 249-56
11. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001; 409: 860-921
12. Tettelin H, Nelson KE, Paulsen IT, et al. Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. Science 2001; 293: 498-506
13. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001; 291: 1304-51
14. Chambers G, Lawrie L, Cash P, et al. Proteomics: a new approach to the study of disease. J Pathol 2000; 192: 280-8
15. Thornton JM. From genome to function. Science 2001; 292: 2095-7
16. Bateman A, Birney E, Cerruti L, et al. The Pfam protein families database. Nucleic Acids Res 2002; 30: 276-80
17. Sigrist CJ, Cerutti L, Hulo N, et al. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 2002; 3: 265-74
18. Henikoff S, Henikoff JG. Protein family classification based on searching a database of blocks. Genomics 1994; 19: 97-107
19. Attwood TK, Bradley P, Flower DR, et al. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 2003; 31: 400-2
20. Ponting CP, Schultz J, Milpetz F, et al. SMART: identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Res 1999; 27: 229-32
21. Servant F, Bru C, Carrere S, et al. ProDom: automated clustering of homologous domains. Brief Bioinform 2002; 3: 246-51
22. Marchler-Bauer A, Anderson JB, DeWeese-Scott C, et al. CDD: a curated entrez database of conserved domain alignments. Nucleic Acids Res 2003; 31: 383-7
23. Mulder NJ, Apweiler R, Attwood TK, et al. The InterPro database 2003 brings increased coverage and new features. Nucleic Acids Res 2003; 31: 315-8
24. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol 1990; 215: 403-10
25. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25: 3389-402
26. Altschul SF. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol 1991; 219: 555-65
27. Dayhoff MO, Schwartz R, Orcutt BC. A model of evolutionary change in proteins. In: Davidoff MO, editor. Atlas of protein sequence and structure. Silver Spring (MD): National Biomedical Research Foundation, 1978: 345-52
28. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992; 89: 10915-9
29. Durbin R, Eddy S, Krogh A, et al. Biological sequence analysis. Cambridge: Cambridge University Press, 1998
30. Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol 1982; 162: 705-8
31. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970; 48: 443-53
32. Smith TF, Waterman MS. Identification of common molecular sub-sequences. J Mol Biol 1981; 147: 195-7
33. Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 1990; 183: 63-98
34. Baldi P, Chauvin Y, Hunkapiller T, et al. Hidden Markov models of biological primary sequence information. Proc Natl Acad Sci U S A 1994; 91: 1059-63
35. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A 1987; 84: 4355-8
36. Jaakkola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. J Comput Biol 2000; 7: 95-114
37. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998; 14: 846-56
38. Madera M, Gough J. A comparison of profile hidden Markov model procedures for remote homology detection. Nucleic Acids Res 2002; 30: 4321-8
39. Park J, Karplus K, Barrett C, et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998; 284: 1201-10
40. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 1997; 28: 405-20
41. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22: 4673-80
42. Vinga S, Almeida J. Alignment-free sequence comparison: a review. Bioinformatics 2003; 19: 513-23
43. Lynch M. Intron evolution as a population-genetic process. Proc Natl Acad Sci U S A 2002; 99: 6118-23
44. Zhang YX, Perry K, Vinci VA, et al. Genome shuffling leads to rapid phenotypic improvement in bacteria. Nature 2002; 415: 644-6
45. Wu CH, Huang H, Yeh LL, et al. Protein family classification and functional annotation. Comput Biol Chem 2003; 27: 37-47
46. Pearson WR. Effective protein sequence comparison. Methods Enzymol 1996; 266: 227-58
47. Pearson WR. Empirical statistical estimates for sequence similarity searches. J Mol Biol 1998; 276: 71-84
48. Blaisdell BE. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci U S A 1986; 83: 5155-9
49. Blaisdell BE. Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a computer-generated model system. J Mol Evol 1989; 29: 538-47
50. Felsenstein J. PHYLIP (Phylogeny Inference Package) [online]. Seattle (WA): Department of Genetics, University of Washington, 1993. Available from URL: http://cmgm.stanford.edu/phylip/#1 [Accessed 2004 Sep 21]
51. Zharkikh AA, Rzhetsky AY. Quick assessment of similarity of two sequences by comparison of their L-tuple frequencies. Biosystems 1993; 30: 93-111
52. Petrilli P. Classification of protein sequences by their dipeptide composition. Comput Appl Biosci 1993; 9: 205-9
53. Solovyev VV, Makarova KS. A novel method of protein sequence classification based on oligopeptide frequency analysis and its application to search for functional sites and to domain localization. Comput Appl Biosci 1993; 9: 17-24
54. Wu TJ, Hsieh YC, Li LA. Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. Biometrics 2001; 57: 441-8
55. Stuart GW, Moffett K, Leader JJ. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. Mol Biol Evol 2002; 19: 554-62
56. Wu TJ, Burke JP, Davison DB. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics 1997; 53: 1431-9
57. Kullback S. Information theory and statistics. New York: Dover, 1968
58. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948; 27: 379-423-656
59. Almeida JS, Carrico JA, Maretzek A, et al. Analysis of genomic sequences by chaos game representation. Bioinformatics 2001; 17: 429-37
60. Almeida JS, Vinga S. Universal sequence map (USM) of arbitrary discrete sequences. BMC Bioinformatics 2002; 3: 6. Epub 2002 Feb 05
61. Li M, Vitanyi P. An introduction to Kolmogorov complexity and its applications. New York: Springer, 1997
62. Baldi P, Brunak S. Bioinformatics: the machine learning approach. Cambridge (MA): MIT Press, 2001
63. Cristianini N, Shawe-Taylor J. An introduction to support vector machines. New York: Cambridge University Press, 2000

64. Vapnik V. The nature of statistical learning theory. New York: Springer-Verlag, 1995

65. Deshpande M, Karypis G. Evaluation of techniques for classifying biological sequences. 6th Pacific-Asia Conference on Knowledge Discovery (PAKDD 2002); 2002 May 6-8; Taipei.

66. Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 2002; 18: 147-59

67. Zavaljevski N, Stevens FJ, Reifman J. Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. Bioinformatics 2002; 18: 689-96

68. Hansen L, Salamon P. Neural network ensembles. IEEE Trans Pattern Anal Mach Intell 1990; 12: 993-1001

69. Krogh A, Vedelsby J. Neural network ensembles, cross validation, and active learning. In: Tesauro G, Touretzky D, Leen T, editors. Advances in neural information processing systems. Vol. 7. Cambridge (MA): MIT Press, 1995: 231-8

70. Opitz D, Maclin R. Popular ensemble methods: an empirical study. J Artif Intell Res 1999; 11: 169-98

71. Wu CH. Gene classification artificial neural system. Methods Enzymol 1996; 266: 71-88

72. Eddy SR. Profile hidden markov models. Bioinformatics 1998; 14: 755-63

73. SAM: sequence alignment and modeling software system. The SAM documentation [technical report no.: UCSC-CRL-95-7]. Santa Cruz (CA): University of California, 1995

74. Sonnhammer EL, Eddy SR, Birney E, et al. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 1998; 26: 320-2

75. Zhang Z, Schaffer AA, Miller W, et al. Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res 1998; 26: 3986-90

76. Okinaka R, Cloud K, Hampton O, et al. Sequence, assembly and analysis of pX01 and pX02. J Appl Microbiol 1999; 87: 261-2

77. Khorana HG. Molecular biology of light transduction by the mammalian photoreceptor, rhodopsin. J Biomol Struct Dyn 2000; 11: 1-16

78. Hwa J, Reeves PJ, Klein-Seetharaman J, et al. Structure and function in rhodopsin: further elucidation of the role of the intradiscal cysteines, Cys-110, -185, and -187, in rhodopsin folding and function. Proc Natl Acad Sci U S A 1999; 96: 1932-5

79. Hwa J, Klein-Seetharaman J, Khorana HG. Structure and function in rhodopsin: mass spectrometric identification of the abnormal intradiscal disulfide bond in misfolded retinitis pigmentosa mutants. Proc Natl Acad Sci U S A 2001; 98: 4872-6

80. Palczewski K, Kumasaka T, Hori T, et al. Crystal structure of rhodopsin: a G protein-coupled receptor. Science 2000; 289: 739-45

81. Altenbach C, Yang K, Farrens DL, et al. Structural features and light-dependent changes in the cytoplasmic interhelical E-F loop region of rhodopsin: a site-directed spin-labeling study. Biochemistry 1996; 35: 12470-8

82. Altenbach C, Cai K, Khorana HG, et al. Structural features and light-dependent changes in the sequence 306-322 extending from helix VII to the palmitoylation sites in rhodopsin: a site-directed spin-labeling study. Biochemistry 1999; 38 (25): 7931-7

83. Altenbach C, Klein-Seetharaman J, Hwa J, et al. Structural features and light-dependent changes in the sequence 59-75 connecting helices I and II in rhodopsin: a site-directed spin-labeling study. Biochemistry 1999; 38 (25): 7945-9

84. Altenbach C, Klein-Seetharaman J, Cai K, et al. Structure and function in rhodopsin: mapping light-dependent changes in distance between residue 316 in helix 8 and residues in the sequence 60-75, covering the cytoplasmic end of helices TM1 and TM2 and their connection loop CL1. Biochemistry 2001; 40 (51): 15493-500

85. Altenbach C, Cai K, Klein-Seetharaman J, et al. Structure and function in rhodopsin: mapping light-dependent changes in distance between residue 65 in helix TM1 and residues in the sequence 306-319 at the cytoplasmic end of helix TM7 and in helix H8. Biochemistry 2001; 40 (51): 15483-92

86. Cai K, Langen R, Hubbell WL, et al. Structure and function in rhodopsin: topology of the C-terminal polypeptide chain in relation to the cytoplasmic loops. Proc Natl Acad Sci U S A 1997; 94: 14267-72

87. Cai K, Klein-Seetharaman J, Farrens D, et al. Single-cysteine substitution mutants at amino acid positions 306-321 in rhodopsin, the sequence between the cytoplasmic end of helix VII and the palmitoylation sites: sulfhydryl reactivity and transducin activation reveal a tertiary structure. Biochemistry 1999; 38: 7925-30

88. Cai K, Klein-Seetharaman J, Altenbach C, et al. Probing the dark state tertiary structure in the cytoplasmic domain of rhodopsin: proximities between amino acids deduced from spontaneous disulfide bond formation between cysteine pairs engineered in cytoplasmic loops 1, 3, and 4. Biochemistry 2001; 40 (42): 12479-85

89. Farrens DL, Altenbach C, Yang K, et al. Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin. Science 1996; 274: 768-70

90. Klein-Seetharaman J, Hwa J, Cai K, et al. Single-cysteine substitution mutants at amino acid positions 55-75, the sequence connecting the cytoplasmic ends of helices I and II in rhodopsin: reactivity of the sulfhydryl groups and their derivatives identifies a tertiary structure that changes upon light-activation. Biochemistry 1999; 38 (25): 7938-44

91. Klein-Seetharaman J, Hwa J, Cai K, et al. Probing the dark state tertiary structure in the cytoplasmic domain of rhodopsin: proximities between amino acids deduced from spontaneous disulfide bond formation between Cys316 and engineered cysteines in cytoplasmic loop 1. Biochemistry 2001; 40 (42): 12472-8

92. Resek JF, Farahbakhsh ZT, Hubbell WL, et al. Formation of the meta II photointermediate is accompanied by conformational changes in the cytoplasmic surface of rhodopsin. Biochemistry 1993; 32: 12025-32

93. Yang K, Farrens DL, Hubbell WL, et al. Structure and function in rhodopsin: single cysteine substitution mutants in the cytoplasmic interhelical E-F loop region show position-specific effects in transducin activation. Biochemistry 1996; 35 (38): 12464-9

94. Cai K, Klein-Seetharaman J, Hwa J, et al. Structure and function in rhodopsin: effects of disulfide cross-links in the cytoplasmic face of rhodopsin on transducin activation and phosphorylation by rhodopsin kinase. Biochemistry 1999; 38 (39): 12893-8

95. Cheng BYM, Carbonell J, Klein-Seetharaman J. Protein classification based on text document classification techniques. Proteins 2004. In press

96. Horn F, Weare J, Beukers MW, et al. GPCRDB: an information system for G protein-coupled receptors. Nucleic Acids Res 1998; 26: 275-9

Correspondence and offprints: Dr *John K. Vries*, Department of Molecular Genetics and Biochemistry, School of Medicine, Center for Computational Biology and Bioinformatics, University of Pittsburgh, 200 Lothrop St, Pittsburgh, PA 15213, USA.
E-mail: vries@ccbb.pitt.edu