Alignment-independent techniques for protein classification

*Matthew N. Davies[1]

Andrew Secker[2]

Alex A. Freitas[2]

Jon Timmis[3]

Edward Clark[3]

Darren R. Flower[1]

[1] The Jenner Institute,
University of Oxford,
Compton, Newbury, Berkshire
RG20 7NN, U.K.

[2]Department of Computing and Centre for BioMedical Informatics
University of Kent,
Canterbury,
Kent CT2 7NF, U.K.

[3]Departments of Computer Science and Electronics
University of York
Heslington
York YO10 5DD, U.K.

Corresponding Author: Matthew Davies, the Jenner Institute, University of Oxford, Compton, Newbury, Berkshire, RG20 7NN. Tel: 0207 631 6842. Email: m.davies@mail.cryst.bbk.ac.uk

Running Title

Alignment-independent techniques

Keywords

Multiple Alignment; Protein Motifs; Alignment independent; Protein Classification; Discrete Form; Sequential form; Local Descriptors; Proteochemometrics;

List of Abbreviations

ACC: AutoCross Co-variance
ANN: Artificial Neural Network
BLOSUM: BLOcks of Amino Acid SUbstitution Matrix
FAO: Food and Agriculture Organisation
GPCR: G protein coupled receptor
HMM: Hidden Markov Model
OET-KNN: Optimised evidence-theoretic K-nearest neighbour
PAM: Point Accepted Mutations
PCA: Principal Component Analysis
PLS: Partial Least Squares
QFC: Quasi-predictor Feature Classifier
SOM: Self Organising Map
SVM: Support Vector Machines
WHO: World Health Organisation

Abstract

Predicting protein structure and function from amino acid sequences is a central aim of bioinformatics. Most bioinformatics analysis uses sequence alignment as the basis by which to measure similarity. However, there is increasing evidence that many protein families are resistant to this straightforward method of comparison. Increasingly, a combination of machine-learning techniques and abstract representations of protein sequences are being used to classify proteins based upon the similarity of their physico-chemical properties rather than scoring sequence alignments. This is particularly effective in protein families that show greater structural conservation but appear to lack conserved sequences. Here we describe the inherent limitations of the alignment-dependent approaches to protein classification and present ‘alignment-free’ representations as a viable and realistic alternative to solve complex problems within bioinformatics.

Introduction

The discovery of new protein sequences is accelerating yet many of these proteins show similarity to existing sequences; the proportion of new but redundant sequences is thus increasing. This may indicate that the global proteome, estimated to be approximately five million sequences, is within a few years of completion (1). In light of this, and the extraordinary productivity of metagenomics (a relatively new field which combines the study of nucleotide sequences with their structure, regulation, and function) (2), efficient bioinformatics analysis of protein sequences to determine structure and function has become an imperative. Computational analysis of the nucleotide and protein sequences of a protein family can reveal the relatedness of its members (3). Protein sequence analysis has evolved into a rational methodology based upon the analysis of multiple sequence alignments, which can determine the similarity between proteins by comparing their primary sequences (4). The technique sorts proteins into recognisable groups using variance maximisation, which exploits the differences between the columns of aligned amino acids (5). Variance maximisation is based on the premise that an arrangement of groups that reflect intrinsic structure or function would feature many columns containing large frequencies of similar amino acids. Such groupings should differ significantly from an average distribution of residues and an optimal arrangement is achieved when there is the maximum variance obtained over all the groups being classified. Protein groups are therefore clustered by their similarity to each other. Sequence comparison takes places in two stages. First, the sequence is aligned with reference to homologous sequences, such that overtly similar regions are brought into register. Secondly, a score is derived from the calculated alignment. The relatedness of two proteins is quantified by calculating the similarity between the aligned sequences. The similarity or mutation matrix describes the probabilities of amino acid mutations for a given period of evolution. This makes the assumption that all mutations occur randomly but are dependent on probabilities defined by properties of the amino acids themselves. This model of evolution computes the probability that the alignment arises by reasons of common ancestry while the product of the database frequencies of the aligned amino acids are used as the probability of alignment by chance. This model assumes Neighbour Independence, each symbol mutates randomly and independently of each other so that the mutation of one amino acid is completely uncoupled from that of its neighbours, positional independence, the probability of mutating from amino acid *i* (A*i*) to amino acid *j* (A*j*) depends only on A*i* and A*j* has the same probability of occurrence in any part of the sequence and historic independence, the model is without memory and the probability of mutation at each site only depends on the present state of the sequence.

The first set of replacement matrices were developed by Dayhoff (6). The replacement rates for the matrices were derived from alignments of protein sequences containing 1572 accepted mutations between 34 superfamilies, all of which had at least 85% homology. Data was collected on point accepted mutations (PAMs) per 100 residues. A 1-PAM mutation matrix describes an amount of evolution which will change, on the average, 1% of the amino acids. No linear relationship is implied between PAM distance and time as evolution rates vary significantly between different species and protein types. The matrices P(0.5), P(1) and P(2.5) (known as the PAM50, PAM100 and PAM250 matrices) are still used to assess the significance of proposed matches between target and database sequences. Commonly used also are the BLOSUM (BLOcks of Amino Acid SUbstitution Matrix) series of matrices, developed by Henikoff and Henikoff (7), which were derived from a database of alignments of protein blocks. BLOSUM matrices improved the accuracy of alignments previously obtained from the PAM matrices (this is demonstrated by the use of the local similarity search method BLAST). While the PAM matrices were estimated from only closely related sequences, the BLOSUM matrices were derived using local, ungapped alignments of distantly related sequences. Matrices of this series are identified by a number after the matrix (e.g. BLOSUM50), which refers to the minimum percentage identity of the blocks of multiple aligned amino acids used to construct the matrix. The BLOSUM matrices often perform better than PAM matrices for local similarity searches, but have not been as widely used in phylogenetics, the study of evolutionary relatedness amongst various organisms. Other matrices developed include Jones et al. (8), an amino acid replacement matrix calculated specifically for membrane spanning segments following the observation that the Dayhoff matrices were biased toward water-soluble globular proteins, and the SCV matrix (9), which focuses on representing solvent accessibility, residue charge and residue volume.

Although the use of mutation matrices remains widespread in bioinformatics, there are certain limitations to the technique. Firstly, the matrices lack the discretion to determine which parts of the sequence are likely to be conserved, typically those relating to structure and function, and which are not. It is also the case that dependencies which relate one amino acid characteristic to the characteristics of its neighbours are not possible to model through this mechanism. The model represents evolution as a Markov process, assuming a consistent rate of evolution where each mutation occurs independently. In actuality, this fails to account for the true effect to multiple substitutions or the inconsistent effect of point mutations on the proteome (10). It is likely that the deterministic forces that have driven the evolution of the genetic code are partly physico-chemical in nature. Natural selection has produced a redundant code that protected the phenotype from errors due to the high frequency of silent point mutations. The structure of the genetic code may also reflect the biosynthetic pathway of amino acid formation (11). However, the relationship between point mutation within the genetic code and amino acid properties is not fully understood. It has been suggested that assuming a steady mutation rate can lead to a suboptimal description of the true phylogeny and that the strategy necessary to optimise a mutation matrix may vary considerably depending on the nature of the proteins or the organism from which they are derived. However, there is evidence that at low divergence between sequences, the genetic code is a better indicator of phylogeny while at high divergence better accuracy is achieved by focussing on amino acid properties (12).

Quantifying the similarity or dissimilarity between two sequences is therefore ambiguous and depends largely on the relative importance that is assigned to a particular region (4). More often the classification of proteins into groups is dependent on the identification of specific motifs within the primary sequence. Commonly these pertain to the protein's active site (13) and these have been used to develop the ‘fingerprints’ by which certain protein groups can be identified (14) or the PROSITE patterns or profiles of protein classification (15).

A more fundamental problem with alignment-based techniques is the intrinsic logistic difficulty encountered when undertaking multiple sequence alignment, particularly when there are several hundred sequences to be aligned (16). There are inherent limitations in alignment-based techniques that may prohibit its application, particularly to protein groups with low sequence similarity. It has been observed that despite continual developments in the technique, there is still a significant drop off in the accuracy of the matrices when there is less than 30% sequence similarity between proteins. Above the ‘twilight’ zone (20-30% homology), accuracies of up to 80% have been reported (17) (although such a figure is very largely on the dataset provided). Within the twilight zone, the accuracy level drops to 65-68%, indicating there is a point where the relationship between similarity score and structural similarity breaks down. This would suggest that there is a point beyond which alignment scores lose the capacity to be a quantitative metric. There are several reasons why this may occur. Alignment-dependent techniques represent the sequences as being essentially linear, ignoring both the physical three-dimensional nature of the protein structure and the dynamic nature of the complex it forms. Hence alignment does not fully account for long distance interactions nor the general fluidity of the actual protein. Another weakness of alignment is that it focuses only on ‘positive’ samples (protein family members) in the dataset without any contribution of ‘negative’ samples (non-members) to the training of the algorithm. It is also possible that the initial alignment of a protein group may introduce a bias into the subsequent alignment of new protein sequences. Lastly, the notion of alignment is based on the premise that protein sequences are contiguous, which is seemingly at odds with processes such as the shuffling of sub-genomic DNA fragments through genetic recombination and exon skipping (4).

Attempts to compensate for the inherent problems of alignment-based techniques have largely concentrated on machine-learning techniques such as Artificial Neural Networks (ANNs) (19), Support Vector Machines (SVMs) (20) or Hidden Markov Models (HMMs) (21). Machine-learning techniques are a subfield of artificial intelligence which are focussed on the design and development of “learning” algorithms. These algorithms are employed in bioinformatics in an attempt to determine more intricate patterns of classification. Machine learning techniques are sometimes based upon the representation of a protein by its dipeptide composition, whereby each of the 400 possible pairs of amino acids is associated with a vector component representing the percentage of the primary sequence consisting of that pair. More recently, however, a different type of analysis has started to emerge, one that is based on the physicochemical properties of proteins and not on any direct or indirect alignment of the sequences. Similar techniques have already been successfully applied to the analysis of genomic sequences, where chaos game representations have shown the capacity to identify similar patterns between related genes of similar species and also shown that sub-sequences of a genome exhibit the main characteristics of the whole genome (22). This means that the Alignment-Free approach is capable of generating a similar score that is not dependent on direct alignment but which is based upon the specific sequential nature of the proteins. Such analysis assumes no homologous relationships between similar sequences and is not reliant on motifs, allowing it to avoid some of the pitfalls of biased sampling. Instead, alignment-free descriptors are designed to extract sequence properties that are shared among functionally similar proteins, making it particularly suitable for the analysis of protein families with low sequence homology. We discuss below Alignment-Free approaches to protein sequence analysis; and how such approaches can be employed to classify and characterise protein families effectively.

Technical Overview of Alignment-Free techniques

Alignment-free techniques represent the physicochemical properties of proteins as numerical values related to their sequence and/or composition. Although some alignment based techniques do incorporate physicochemical properties as part of their methodology (23) This technique uses amino acid indices, which measure the relative values of a certain property such as hydrophobicity, polar values or steric values (24). Each index only measures a single property, and do not describe fully all the attributes of a residue. With so much redundancy contained within available indices, the physicochemical properties of each side-chain must be summarised into a few key values that sufficiently characterise each residue. Sequence similarity can be assessed by combining many amino acid properties into a substitution or mutation matrix. Large numbers of amino acid indices and mutation matrices are contained in the Amino Acid Index (24).

There are two ways in which the alignment-free techniques can represent a protein sequence, the discrete form and the sequential. The discrete form is based on a protein's amino acid composition, consisting of the frequencies of the 20 different amino acids in the sequence. The sequential form also accounts for the order of amino acids in a sequence. While the discrete form is more manageable than the sequential, it possesses certain limitations. In the discrete form, the sequence order and length are lost. While composition is important, to exploit fully all information implicit within a protein sequence, we must incorporate information from the explicit order of amino acids. However, as protein sequences have variable lengths, normalisation of resulting derived data is necessary.

A simple example of an alignment-free approach is the decision tree induction algorithm (25), which is applied to attributes representing dipeptide compositions. More precisely, each protein is represented by a vector of 20 values, representing the frequency of the 20 amino acids in the sequence. The C4.5 algorithm is used to create subsets from the training set based on amino acid compositions (26). The choice of when to split is made by selecting the amino acid composition feature which best discriminates the classes to be predicted. The division continues until the stopping criteria are satisfied and a successful classification algorithm is generated.

Optimised evidence-theoretic K-nearest neighbour (OET-KNN) is a classification method somewhat similar to an SVM (27). It has been used to classify proteins represented by amino acid composition. The basic idea of a K-NN algorithm is that, for each unknown protein sequence, the system computes a measure of the similarity between that sequence and each of the known sequences in the training set. Then the system selects the K nearest training sequences, and the class of the majority of those sequences is transferred to the class of the unknown sequence. The technique has been applied successfully to various transmembrane protein families.

One of the first attempts are representing proteins sequences based on amino acid composition was the PropSearch program (http://abcis.cbs.cnrs.fr/propsearch) (28). The program described the protein sequence as having 144 properties including amino acid composition, hydrophobicity, charge and 'doublet' groups showing the incidence of pairing of amino acids with similar or different physical properties within the sequence. This provides a measure of the variation of the physical properties within the protein. Multiple sequences could be merged into an average sequence that reflected the properties of a specific protein family. Weights were optimised between protein subtypes to allow maximal discrimination between protein subtypes.

A similar approach is used for the Local Protein Sequence descriptors method, a discrete form of analysis whereby the protein sequence is divided into ten regions, the length and composition of which is dependent on the total size of the protein (29-30). Three descriptors - composition (C), transition (T) and distribution (D) - represent the properties of each local region. C represents the percentage of frequencies of residues with a specified property that occur in the local region, T represents the percentage frequency with which the specified property changes within the region while D characterises the distribution of the property along the entire region by measuring the location at which the first, 25%, 50%, 75% and 100% of the residues with the property occur. Descriptors for all local regions over the whole protein are combined to give a general representation of the sequence. Key to this analysis is the grouping of the twenty amino acids used, because a descriptor value refers to a group of similar amino acids rather than being associated with a single residue. As there is no conventional means of grouping the standard twenty amino acids, Cui *et al.* (30) put them into the three categories of hydrophobic (CVLIMFW), neutral (GASTPHY) and polar (RKEDQN). However, many side chains have properties pertaining to more than one group. For instance, tyrosine is characterised as neutral but has an aromatic ring and a hydroxyl group: one suggests hydrophobicity, the other polarity. Different amino acid groupings would be expected to produce different results. It is also difficult to anticipate the optimal number and distribution of groups. Too many groups might better represent the residues but could overly complicate the descriptors. Optimising the amino acid groupings using data mining techniques (31) showed clearly that the same groupings to appear, in particular the grouping of cysteine on its own, indicating a unique property of the residue.

Sequential form analysis focuses on finding periodicity within protein sequences. The calculations are performed for a lag value of 1, e.g. adjacent residues, then for a lag value of 2, 3 and so on. One simple way to incorporate sequential information is by describing the balance of hydrophobic and hydrophilic properties with a protein. A secondary structural property, such as an ideal alpha helix, will naturally balance hydrophobic and hydrophilic residues, creating an amphiphilic sequence. The amphiphilic properties of different proteins will vary. The amphiphilic pseudo amino acid composition (Am-Pse-

AA composition) exploits such variations to differentiate protein types (19). It incorporates the classic amino acid composition and supplements this with the amino acid sequence correlation along the length of the protein.

It is, however, possible to incorporate far more than just amphiphilic values into a sequential representation. Kim *et al.* (32-33) classify sequences in a ‘feature space’ and create discrimination functions that put sequences into specific categories. The feature space uses statistical measures of physico-chemical properties and then uses a linear discriminant function to classify protein sequences. The Quasi-predictor Feature Classifier (QFC) algorithm was designed to statistically characterize the differentiating features of the physico-chemical properties of protein sequences using heuristic data reduction principles. The amino acid properties examined were the GES hydropathy, the Kyte-Doolittle index, polarity, isoelectric point, molecular weight, solubility and the alpha helix index. The values were also normalized using the Sliding Windows Recogniser with the separate values being summed over a window. Window lengths of 13-16 amino acids were shown to be more effective than lengths of 32 and 64 amino acids. However, the QFC algorithm was shown to have a higher false positive rate than most motif-based techniques, suggesting that the technique would benefit from another stage of filtering.

Proteochemometrics is a technique whereby 26 separate physicochemical properties of the protein are used to calculate five empirical ‘z’ values for all 20 amino acids (34-36). The z1 value accounts for the amino acid’s lipophilicity while the z2 value summarises the residue’s steric bulk/polarisability. Polarity of the amino acid is described by the z3 value while the z4-5 values represent the electronic effects. These five values are calculated for each amino acid in the sequence, generating a matrix that provides a purely numerical description of the protein’s character. AutoCross Co-variance (ACC)(37) is used to normalize the uneven size of the z matrices and then Principal Component Analysis (PCA) and Partial Least Squares (PLS) are carried out in order to provide a classification system for proteins. As the proteins used in the study had different lengths, an auto cross covariance (ACC) transformation was used to transform them to a uniform length. ACC calculates lag values between the residues positions within a z value (Auto contribution) and between z values (Cross contribution). The results of these transformations were new uniform sets of variables for each protein. Key to the ACC approach is balancing the two factors of maximum lag, L, and the degree of normalization, p. Optimal parameter values can be determined by trying various combinations and then assessing their total classification accuracy. However, further research by Secker *et al.* (38) has suggested that using an arithmetic mean over each attribute for each protein could be an effective means to normalise the data while vastly reducing the processing time and storage requirements necessary to implement the Auto Cross Covariance approach. It should be noted that this approach therefore applies a discrete representation to the data rather than a sequential one. Proteochemometrics is also used in the application of Self-Organising Maps (SOMs) (39). SOMs are Artificial Neural Networks (ANNs) that perform unsupervised learning (in this case, clustering) to determine one protein family from another. Unlike PCA, which relies upon establishing linear relations, an SOM can accommodate non-linear relations into its algorithm. In order to make a family map to determine a certain protein subtype, it is necessary to determine a family area that contains the most frequent "activator" family samples. A feature map developed for protein classification must also include an objective border between clusters that clearly distinguishes each family.

Application of alignment-free techniques

There are several instances where the application of alignment-free techniques have been proven to be effective where alignment based methods have not been. In such cases where there is a sequence similarity <30% between two sequences (sometimes described as the 'twilight zone'), it is extremely difficult for them to be correctly aligned. Protein families that have great structural and/or functional homology but a low degree of sequence similarity are therefore very difficult to classify. An example of this is G protein-coupled receptors (GPCRs), which play an important role in many physiological systems by transducing an extracellular signal into an intracellular response.

There is considerable interest in developing an algorithm that could effectively predict the function of a GPCR from its primary sequence. Such an algorithm is useful not only in identifying novel GPCR sequences but in characterising the interrelationships between known GPCRs. An extremely heterogeneous set of molecules can act as GPCR ligands including ions, hormones, neurotransmitters, peptides and proteins. The GPCRs are a common target for therapeutic drugs and approximately 50% of all marketed drugs target GPCRs (40-41). In spite of their functional and sequence diversity, there are certain structural features common to all GPCRs. All GPCRs contain seven highly conserved transmembrane segments. The transmembrane segments form seven α-helices in a flattened two-layer structure known as the transmembrane bundle, a structure seen in all GPCRs (42). The GPCRs shows a far greater conservation with regard to the three-dimensional structure than to the primary sequence. The diversity of the GPCRs means it is difficult to develop a comprehensive classification system for all of the GPCR subtypes (43). Although the GPCRs can be grouped by structure but only subsets can efficiently be grouped by sequence. Previous attempts at predicting the function of a GPCR from its primary sequence, and therefore its position within a given hierarchical system, have included motif-based classification tools (44-45) and machine learning methods such as Hidden Markov Models (46). The majority of predictive techniques, however, have used Support Vector Machines (SVMs) (47), machine-learning algorithms based on statistical learning theory. Although SVMs are more commonly used to solve 2-class problems, this technique can be applied to the classification of GPCR data by successively trying to classify one class against all others. Several publicly available SVM-based GPCR classifiers exist including PRED-GPCR (48-49), GPCRPred (50) and, GPCRsclass (51), which concentrates on the Class A aminergic receptor subfamily. In the first round of analysis, an SVM is generated to distinguish amines from all other GPCRs.

An alignment-free approach to GPCR classification was developed using techniques drawn from data mining and proteochemometrics from a dataset of over 8,000 sequences. The predictive algorithm was developed based upon the simplest reasonable numerical representation of the protein's proteochemometric z values (34-36). A selective top-down approach was developed which used a classifier to assign sequences to subdivisions within the GPCR hierarchy (52). The predictive performance of the algorithm was assessed against several standard data mining classifiers and further validated against Support Vector Machine-based GPCR prediction servers. The selective top-down approach was shown to have significantly higher accuracy than standard data mining methods in almost all cases. This suggests an Alignment-Free representation is the best approach to this particular set of proteins.

Another example of where the application of an Alignment-Free technique can be effective is in the detection of allergens. Allergic reactions are dependent on a series of intrinsic and extrinsic factors that control both the development and triggering of the condition (53). They are often caused by the type I hyperreactive reaction, induced by antigens that elicit specific IgE antibodies or from cross-reactivity between similar allergens. The increasing use of modified proteins in food has lead to an increasing concern about the identification of allergenic proteins. The World Health Organisation (WHO) and Food and Agriculture Organisation (FAO) have defined an allergen as any protein that shares a contiguous sequence of at least six residues with a known allergen or has a homology with it of >35% over a region of 80 residues. This definition has been heavily criticised since it produces so many false positive results. Previous approaches to IgE prediction (IgE is a type of antibody only found in mammals that plays an important role in allergies) have used motif-based approaches (the MEME/MAST program) (54), a K-nearest neighbour classifier (55) and similarity searches against IgE epitope/epitope profiles (56). However, none of the techniques developed so far have produced a reliable predictor of IgE epitopes, partly due to the limited number of known epitopes available. Many allergens appear to cluster in a limited number of protein families but very few families seem to lack members with allergenic properties (57). There is considerable interest in designing allergens that can be modified to address specific cells of the immune system and hence to dictate the extent of the immune response created (58-59) although this is not possible without a strategy to determine proteins with allergenic properties from those with non-allergenic. AllerTool (29) uses a SVM that encodes descriptors derived from amino acid composition and showed a predictive sensitivity and specificity of 86% in determining allergens from non-allergens.

Conclusion

Alignment based techniques have many advantages, particularly that they are readily understood. However, as with all protein classification approaches, alignment based approaches exhibit many limitations. There are classification problems in bioinformatics that are too subtle for conventional alignment-dependent analysis. It may be that alignment-based analysis cannot properly capture the uneven variance of genetic recombination or the subtle physico-chemical properties that underlying protein structure and folding. Alignment-free techniques also presents certain limitations as, following classification, it is difficult to establish what the properties are of protein sequence assigned to a specific group. Nonetheless, alignment-free analysis offers greater potential for pattern recognition within protein classes and can offer an accurate means of measuring protein similarity. Alignment-free approaches offer a valid alternative to alignment-based techniques and have a wide range of applications in the field of bioinformatics.

Acknowledgements

The authors gratefully acknowledge the ESPRC grant EP/D501377/1.

Conflict of Interest

The authors are unaware of any conflict of interest relating to this manuscript

References

1. Perez-Iratxeta C, Palidwor G, Andrade-Navarro MA. Towards completion of the Earth's proteome. *EMBO Rep*. **2007**; 8(12):1135-41.

2. Raes J, Foerstner KU, Bork P. 2007. Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol* **2007;** 10(5):490-8

3. Fuchs R. From sequence to biology: the impact on bioinformatics. *Bioinformatics* **2002**; 18:505-506.

4. Vinga S, Almeida J. Alignment-free sequence comparison - a review. *Bioinformatics* **2003**; 19:513-23.

5. Wrabl JO, Grishin NV. Grouping of amino acid types and extraction of amino acid properties from multiple sequence alignments using variance maximization. *Proteins* **2005**; 61:523-34.

6. Dayhoff.MO, Schwartz RM, Orcutt, BC. In Dayhoff, MO (ed.). Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, DC **1978**; 5:345-352.

7. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*. **1992**; 89:10915-9.

8. Jones DT, Taylor WR, Thornton JM. A mutation data matrix for transmembrane proteins. *FEBS Lett*. **1994**; 339:269-75.

9. Goodarzi H, Katanforoush A, Torabi N, Najafabadi HS. Solvent accessibility, residue charge and residue volume, the three ingredients of a robust amino acid substitution matrix. *J Theor Biol.* **2007**; 245:715-25.

10. Keane TM, Creevey CJ, Pentony MM, Naughton TJ, Mclnerney JO. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol*. **2006**; 24:29.

11. Goodarzi H, Najafabadi HS, Hassani K, Nejad HA, Torabi N. On the optimality of the genetic code, with the consideration of coevolution theory by comparison of prominent cost measure matrices. *J Theor Biol.* **2005**; 235:318-25.

12. Tomii K, Kanehisa M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng*. **1996**; 9:27-36.

13. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ. 2006. The PROSITE database. *Nucleic Acids Res.* **2006**; 34:D227-30.

14. Attwood TK. The PRINTS database: a resource for identification of protein families. *Brief Bioinform*. **2002**; 3:252-63.

15. Hulo N., Bairoch A., Bulliard V., Cerutti L., De Castro E., Langendijk-Genevaux P.S., Pagni M., Sigrist C.J.A. The PROSITE database. *Nucleic Acids Res*. **2006**; 34:D227–D230.

16. Pham TD, Zuegg J. A probabilistic measure for alignment-free sequence comparison. *Bioinformatics* **2004**; 20:3455-61.

17. Kurgan L, Chen K. 2007. Prediction of protein structural class for the twilight zone sequences. *Biochem Biophys Res Commun*. **2007**; 357:453-60.

18. Murvai J, Vlahovicek K, Szepesvari C, Pongor S. Prediction of protein functional domains from sequences using artificial neural networks. *Genome Res* **2001**; 11:1410–1417.

19. Melvin I, Ie E, Kuang R, Weston J, Stafford WN, Leslie C. SVM-Fold: a tool for discriminative multi-class protein fold and superfamily recognition. *BMC Bioinformatics* **2007**; 228:S2.

20. Srivastava PK, Desai DK, Nandi S, Lynn AM. HMM-ModE--improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. *BMC Bioinformatics* **2007**; 278:104.

21. Almeida JS, Vinga S. 2006. Computing distribution of scale independent motifs in biological sequences. *Algorithms Mol Biol.* **2006**; 1:18.

22. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**; 21:10-9.

23. Frimurera TM, Ulvena T, Ellinga CE, Gerlacha L-O, Kostenisa E, Högberg T. A physicogenetic method to assign ligand-binding relationships between 7TM receptors. *Bioorganic & Medicinal Chemistry Letters* **2005**; 15(16):3707-3712.

24. Kawashima S, Kanehisa M. AAindex: amino acid index database. *Nucleic Acids Res*. **2000**; 28:374.

25. Huang Y, Cai J, Ji L, Yanda L. Classifying G-protein coupled receptors with bagging classification tree. *Computational Biology and Chemistry* **2004**; 28:275-280.

26. Quinlan JR. C4.5: programs for machine learning, Morgan Kaufmann, San Francisco. **1993**.

27. Shen HB, Chou KC. Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. *Biochem Biophys Res Commun*. **2005**; 337:752-6.

28. Hobohm U, Sander C. A sequence property approach to searching protein database. *J. Mol. Biol.* **1995**; 251:390-399.

29. Zhang ZH, Koh JL, Zhang GL, Choo KH, Tammi MT, Tong JC. AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in proteins. *Bioinformatics* **2007**; 23:504-6.

30. Cui J, Han LY, Li H, Ung CY, Tang ZQ, Zheng CJ, Cao ZW, Chen YZ. Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. *Mol. Immunol.* **2007**; 44:514-20.

31. Davies, MN, Secker A, Freitas AA, Clark E, Timmis J, Flower DR. 2007. Optimizing amino acid groupings for GPCR Classification. *Bioinformatics*. Under Review

32. de Trad CH, Fang Q, Cosic I. Protein sequences comparison based on the wavelet transform approach. *Protein Eng.* **2002**; 15:193-203.

33. Kim J, Moriyama EN, Warr CG, Clyne PJ, Carlson JR. Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties. *Bioinformatics* **2000**; 16:767-75.

34. Mathura VS, Kolippakkam D. APDbase: Amino acid Physico-chemical properties Database. *Bioinformation* **2005**; 121:2-4.

35. Strömbergsson H, Kryshtafovych A, Prusis P, Fidelis K, Wikberg JE, Komorowski J, Hvidsten TR. Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures. *Proteins* **2006**; 65:568-79.

36. Lapinsh M, Gutcaits A, Prusis P, Post C, Lundstedt T, Wikberg JES. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Sci*. **2002**; 11:795-805.

37. Wold S, Jonsson J, Sjöström M, Sandberg M, Rännar S. DNA and peptide sequences and chemical processes mutlivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Anal. Chim. Acta*, **1993**; 277:239–253.

38. Secker A, Davies MN, Freitas AA, Timmis J, Mendao M, Flower, DR. An Experimental Comparison of Classification Algorithms for the Hierarchical Prediction of Protein Function. Expert Update (magazine of the British Computer Society's Specialist Group on Artificial Intelligence) **2007**; 9(3):17-22.

39. Otaki JM, Mori A, Itoh Y, Nakayama T, Yamamoto H. Alignment-free classification of G-protein-coupled receptors using self-organizing maps. *J. Chem. Inf. Model.* **2006**; 46:1479-90.

40. Flower DR. Modelling G-protein-coupled receptors for drug design. *Biochim Biophys Acta.* **1999**; 1422:207-234.

41. Klabunde T, Hessler G. Drug design strategies for targeting G-protein-coupled receptors. *ChemBioChem* **2002**; 3:928–944.

42. Milligan G. G-protein-coupled receptor heterodimers: pharmacology, function and relevance to drug discovery. *Drug Discov Today* **2006**; 11:541-9.

43. Davies MN, Gloriam DE, Secker A, Freitas AA, Clark E, Timmis J, Flower DR. Proteomic applications of automated GPCR classification. *Proteomics* **2007**; 7(16):2800-14.

44. Attwood TK. A compendium of specific motifs for diagnosing GPCR subtypes. *Pharmacol Sci* **2001**; 22:162-5.

45. Flower DR, Attwood TK. Integrative bioinformatics for functional genome annotation: trawling for G protein-coupled receptors. *Semin Cell Dev Biol.* **2004**; 15:693-701.

46. Wistrand M, Kall L, Sonnhammer EL. A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Protein Sci*. **2006**; 15:509-21.

47. Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* **2002**; 18:147-159.

48. Papasaikas PK, Bagos PG, Litou ZI, Promponas VJ, Hamodrakas SJ. PRED-GPCR: GPCR recognition and family classification server. *Nucleic Acids Res*., **2004**; 32:W380-2.

49. Guo YZ, Li ML, Wang KL, Wen ZN, Lu MC, Liu LX, Lin J. Fast fourier transform-based support vector machine for prediction of G-protein coupled receptor subfamilies *Acta Biochim Biophys Sin (Shanghai)* **2005**; 37:759-66.

50. Bhasin M, Raghava GP. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res*. **2004**; 32:W383-9.

51. Bhasin M, Raghava GP. GPCRsclass: a web tool for the classification of amine type of G protein-coupled receptors *Nucleic Acids Res*. **2005**; 33:W143-7.

52. Davies MN, Secker A, Freitas AA, Mendao M, Timmis J, Flower DR. On the hierarchical classification of G Protein Coupled Receptors. *Bioinformatics* **2007**. 23:3113-3118.

53. Crameri R, Rhyner C. Novel vaccines and adjuvants for allergen-specific immunotherapy. *Curr Opin Immunol.* **2006**; 18:761-8.

54. Stadler MB, Stadler BM. Allergenicity prediction by protein sequence. *FASEB J.* **2003**; 17:1141-3.

55. Soeria-Atmadja D, Zorzet A, Gustafsson MG, Hammerling U. Statistical evaluation of local alignment features predicting allergenicity using supervised classification algorithms. *Int Arch Allergy Immunol.* **2004**; 133:101-12.

56. Ivanciuc O, Schein CH, Braun W. SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res.* **2006**; 31:359-62.

57. Soeria-Atmadja D, Lundell T, Gustafsson MG, Hammerling U. Computational detection of allergenic proteins attains a new level of accuracy with in silico variable-length peptide extraction and machine learning. *Nucleic Acids Res*. **2006**; 34:3779-93.

58. Linhart B, Valenta R. Molecular design of allergy vaccines. *Curr Opin Immunol.* **2005**; 17:646-55.

59. Bousquet J, Lockey R, Malling HJ. Allergen immunotherapy: therapeutic vaccines for allergic diseases. A WHO position paper. *J Allergy Clin Immunol.* **1998**; 102:558-62.