Pacific Symposium on Biocomputing 7:625-636 (2002)

USING EVOLUTIONARY METHODS TO STUDY G-PROTEIN COUPLED RECEPTORS

ORKUN SOYER[*], MATTHEW W. DIMMIC[#], RICHARD R. NEUBIG[&]
RICHARD A. GOLDSTEIN[*#]
[*]*Department of Chemistry,* [#]*Biophysics Research Division,*[&]*Department of Pharmacology*
University of Michigan, Ann Arbor, MI 48109-1055

A novel method to analyze evolutionary change is presented and its application to the analysis of sequence data is discussed. The investigated method uses phylogenetic trees of related proteins with an evolutionary model in order to gain insight about protein structure and function. The evolutionary model, based on amino acid substitutions, contains adjustable parameters related to amino acid and sequence properties. A maximum likelihood approach is used with a phylogenetic tree to optimize these parameters. The model is applied to a set of Muscarinic receptors, members of the G-protein coupled receptor family. Here we show that the optimized parameters of the model are able to highlight the general structural features of these receptors.

1 Introduction

One of the current main challenges in life sciences is understanding the machinery of biological systems. In the core of this machinery lies proteins; our understanding of biological systems is bound to our knowledge of protein structure and function. There has been increasing interest in obtaining information on structure and function from the rapidly-increasing databases of protein sequences, often through the comparison of related sequences. Despite these efforts, there are currently no generally-applicable methods to derive detailed structural and functional information from such investigations.

Like all biological entities, proteins are a result of evolution. They have developed their current structure and function under the influence of billions of years of selective pressure. Analyzing a family of proteins from different species can unravel information regarding this selective pressure. Such a study might allow one to detect and interpret the selective pressures that have acted on these proteins, providing insight about their structure and function.

In order to be fruitful, such examination requires careful choice of the protein family to be examined and the model of evolution to be used. Here we present the application of an evolutionary model based on amino acid substitution to a family of G-Protein coupled receptors (GPCRs). Located in the cell membrane, these receptors activate the associated G-protein bound to their intracellular part upon binding a ligand on their extracellular side. GPCRs constitute one of the largest protein families, making up 3% to 5% of the coding regions in the human genome[1]. They are associated with many signaling pathways in different cells ranging from neurons to muscle cells, and are targeted by more than 50% of all drugs[1]. In addition to their biological significance, the high structural and functional resemblance among family members is another reason to use GPCRs in evolutionary studies. Throughout the family the general topology of seven transmembrane helices connected by extra- and intracellular loops is highly conserved (see Fig.1). Despite this conserved general topology, GPCRs are able to achieve different functions by coupling to different ligands and/or G-proteins. The sequence similarity among family members is low due to varying composition and length of loops, while highly conserved transmembrane regions allow for reliable sequence alignments. Currently there are more than 3500 known GPCR sequences with only one crystal structure solved, that of bovine rhodopsin (PSB code 1F88)[2].

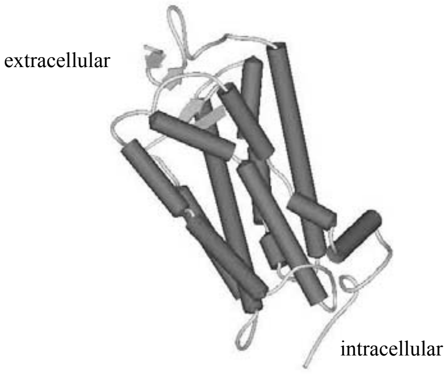

Figure 1: Representation of the crystal structure of bovine rhodopsin. The helical parts, including the seven transmembrane helices, are shown as red cylinders

.

All these properties make GPCRs a good candidate for sequence based studies and there are many examples of such studies in the literature. Most important of these are techniques based on pattern recognition[3] and correlation analysis[4]. These analyses are mainly focused on defining key residues responsible for ligand and/or G protein binding. While these methods have provided important information about specific residues, they are unable to generate more general information about how the observed protein properties are determined by the sequence. In addition, the correlation analysis explicitly neglects evolutionary relationships between the proteins, making them susceptible to misinterpreting correlations induced by the phylogenetic relationships.

2 Model

Evolution proceeds from the fixation of errors occurring during DNA replication. This is generally represented by a substitution matrix, encoding the relative rate at which every possible amino acid or nucleotide substitution occurs on the evolutionary timescale[5,6]. These matrices generally assume that all locations in all proteins can be represented by the same model. Despite the success of these matrices, there are shortcomings both in their creation and use. Derived from a particular set of proteins these matrices might not be able to mimic the substitutions in a different protein family. Their use for different locations of a given protein is also questionable. Given the structural and functional constraints on a protein both the rate and nature of substitutions among different locations should vary.

There have been attempts to incorporate absolute rate heterogeneity among locations by having the substitution rates multiplied by a site-specific scaling factor[7]. While these models are better able to represent biological data, they cannot account for qualitative variations in the type of selection pressure at various locations. Other models have been developed that allow for different locations to be under different types of selective pressure, either due to differences in local structure[8-10] or by allowing every location to be described by a different model[11]. The former method ignores differences in selective pressure due to other factors than local structure, while the latter is limited by the amount of available data. We (and others) have developed methods that allow for variation at different locations by postulating that there are a number of different types of locations, each describable with a specific substitution model, where the assignment of locations to different types is not known a priori[12-15]. In our model this is achieved by using the notions of amino acid fitness and site classes.

The basis of our evolutionary model based on amino acid substitution has been described previously[14,15]. In brief, we encompass the distribution of selective pressures at different locations in the protein by assuming that each location under consideration can be described by one of a number of possible site classes; each with its own set of parameters defining the substitution rates. The model does not assign locations to site classes, instead we define an unknown prior probability $P(k)$, that any given location belongs to site class k. As all locations must belong to a site class, $\Sigma_k P(k)=1$. We also imagine that there is a relative fitness $F_k(A_i)$ of amino acid A_i for any location described by a particular site class k. For example, at the core of a protein we expect a hydrophobic amino acid to have a high fitness value, however we do not impose such expectations on the model *a priori*. We further argue that the probability of substitution between two amino acids should directly depend on the change in fitness values resulting from such substitution. Thus for each site class we define a matrix for all possible substitutions based on the fitness values. Our particular model uses a function, composed of Gaussian and sigmoidal distributions, to calculate the substitution matrix for a small interval of time:

$$Q(k)_{ij} = v_k e^{-\lambda_k (\Delta F_k)^2} \frac{\beta_k e^{\Delta F_k / 2}}{\beta_k e^{|\Delta F_k / 2|} + e^{-|\Delta F_k / 2|}} \qquad (1)$$

where v_k is the substitution rate for site class k, λ and β are parameters of the function, and

$$\Delta F_k = F_k(A_j) - F_k(A_i) \qquad (2)$$

The use of this so-called "gaumoidal" function allows us simulate two different ideas about the process of evolution simultaneously. For small values of λ, the above function will approach a sigmoidal function where substitutions are accepted with v_k if favorable and tolerated with a decaying probability if unfavorable. For large values of λ, the function will approach a Gaussian distribution where conservative substitutions are favored.

To determine the substitution matrix *M*, representing the possible substitutions from amino acid A_i to A_j for any particular amount of evolutionary time *t*, the *Q* matrix is exponentiated:

$$M(k)(t) = e^{tQ(k)} \qquad (3)$$

At each location *l*, the likelihood L_l can be calculated as the probability of data given the model's parameters θ and the evolutionary tree topology and branch lengths. Since each location can be represented by any of the site classes and each site class has distinct parameters θ_k we have to sum over all possible site classes to calculate this likelihood:

$$L_l = \sum_k P_l(Data \mid \theta_k, T) P(k) \qquad (4)$$

with the likelihood for the tree calculated as the product of the likelihood at each location.

The parameters of the model can be optimized using a maximum likelihood approach on a given tree. To summarize there are 23 parameters per site class: 20 amino acid fitness values F_k (with one held constant since the fitness values are relative), substitution rate v_k, gaumoidal function parameters λ and β, and the prior probability for that site class *P(k)*. One of the site class prior probabilities will depend on the others since all priors must add to 1. The initial values for substitution rates are derived from a gamma distribution as described by Yang[16], while the other parameters are set to arbitrary initial values.

While we do not know to which site class a location belongs *a priori*, we can calculate *a posteriori* probabilities. The conditional probability that a location *l* belong to site class *k* is given by:

$$P(k \mid Data_l) = \frac{P(Data_l \mid \theta_k)P(k)}{\sum_k P(Data_l \mid \theta_k)P(k)} \qquad (5)$$

This equation allows us to group locations in the protein that are under similar selective pressure; the parameter values give us insight into the nature of the selective pressure at these locations.

3 Data and Methods

The model explained above is used to predict the structural and functional properties of a subfamily of the GPCR family. The selected subfamily was that of Muscarinic Receptors. These receptors are activated upon binding of acetylcholine and initiate a set of diverse events in the cell through the associated G protein. There are five known types of Muscarinic Receptors that couple to two different G proteins. The data set contained twenty-two receptor sequences from eight different species, representing all five types of Muscarinic Receptors. The sequences and the corresponding multiple alignment of length 530 are obtained from GPCRdb[17].

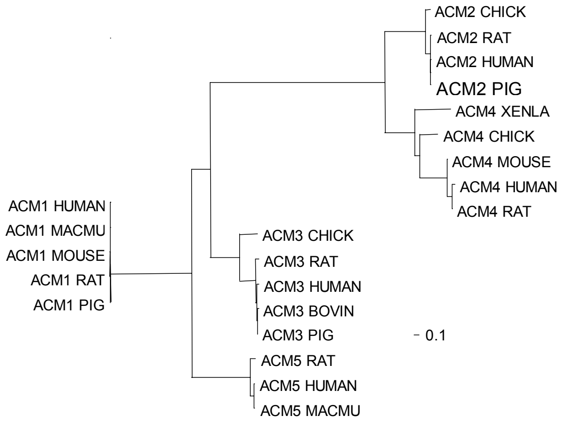

Figure 2: Unrooted phylogenetic tree of ACM receptors. Branch lengths are scaled to number of substitutions along each branch, with the given scale representing 1 substitution per 10 sites.

In order to optimize the parameters of the model a phylogenetic tree of the selected proteins, shown in Fig. 2, was created using PROTML[18]. This software uses a maximum likelihood approach to search for the most likely tree topology. We used this program with the default settings, which use automatic search and the JTT matrix of Jones et. al.[6]. The branch lengths of the resulting tree were further optimized using PAML[19], along with the alpha parameter of gamma distribution used in determining rate variation among site classes.

We optimized our model using the tree with optimized branch lengths for increasing number of site classes. For each run the initial rates for each site class are derived from the gamma distribution using the alpha parameter optimized with PAML. The software optimizes all the parameters for each site class and calculates the posterior probability of each location being represented by any site class.

4 Results

The results of this study consist of optimized fitness values for each site class and the posterior probabilities for each alignment position. We ran the program with two to eight site classes. The resulting optimal parameters are listed in Table 1. In order to interpret the resulting fitness functions, we determined the correlation coefficient between the values of $F_k(A_i)$ and 145 selected amino acid indices from the AAindex database[20]; the two most highly-correlated indices are also listed in Table 1.

The posterior assignment of site classes was mapped onto the structure of the Muscarinic type 3 receptor from human (ACM3_Human). Fig 3 shows the results for the optimization of two site classes. The same plot also contains the hydropathy plot for this receptor. Hydropathy plots are generally used to detect transmembrane regions of membrane proteins. These plots show the seven transmembrane helices of GPCRs clearly and are generally used to predict their sequence location. The correlation between the posterior assignments into the two site classes and the hydropathy plot show that our model assigns the residues into site classes according to their location. Almost all non-transmembrane residues are assigned to site class 2, while almost all transmembrane residues are assigned to site class 1. The results also show that certain conserved non-transmembrane residues such as residues from 2nd and 3rd intracellular loops are also assigned to site class 1 along with most of the transmembrane residues.

The posterior information for larger numbers of site classes is harder to interpret with such simple plots. To see the results from runs with more site classes we converted the posterior information into color strips, where each color represents an assignment of the location to a given site class. Fig.4 shows these strips for different numbers of site classes.

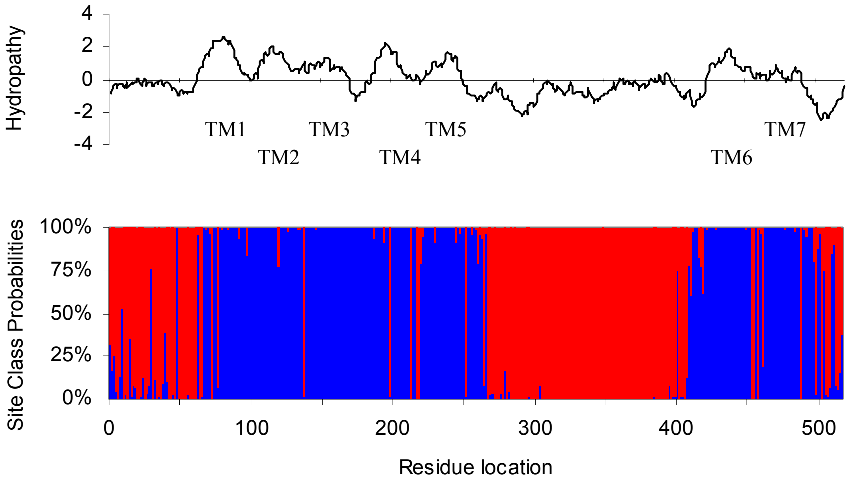

Figure 3: Correlation between hydropathy plot and posterior probabilities for ACM3_Human. Top plot: Kyte-Dolittle hydropathy index. Bottom plot: Relative probability that a location is assigned to site class 1 (blue) or 2 (red). Putative transmembrane helices, identified as the peaks on the hydropathy plot, are marked.

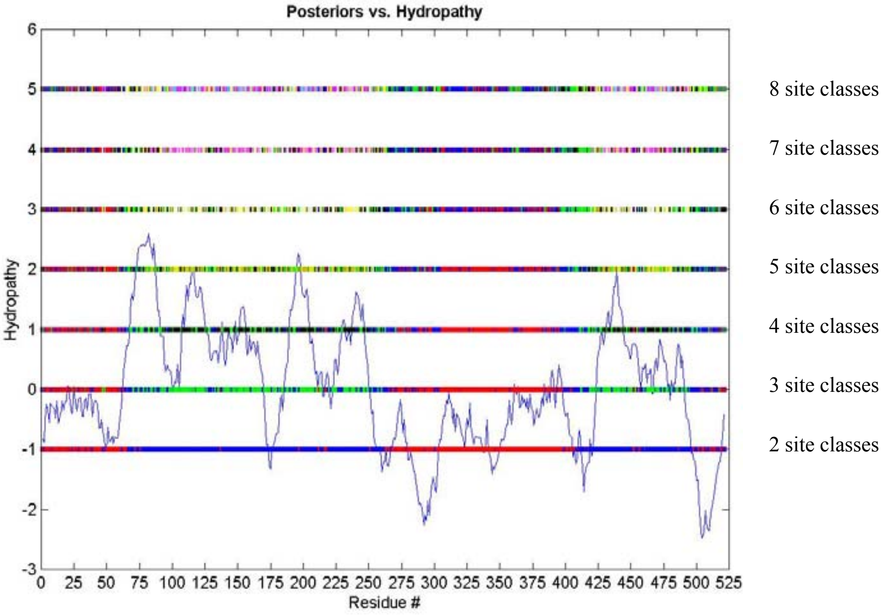

Figure 4: Color strips from posterior probabilities. The strips are matched to the hydropathy plot. Different colors represent different site classes.

As with two site classes, these color strips also indicate a distinct classification of certain residues to different site classes. This classification seems to follow the general topology of GPCRs, which is seven transmembrane helices connected with intra- and extracellular loops. Most of the transmembrane residues fall into same site class regardless of the number of site classes in the model. The residues from the non-transmembrane regions distribute themselves among a set of site classes, generally not including the site classes from the transmembrane helices. As the number of site classes is increased, certain non-transmembrane residues are assigned to site classes characteristic of the transmembrane residues, possibly involving locations where hydrophobicity is important for structural or functional reasons.

# of site	**site**	**parameters**			**1st correlation**		**2nd correlation**	
classes	**class #**	**subst. rate**	**lambda**	**beta**	**Corr. Coeff.**	**Property Code**	**Corr. Coeff.**	**Property Code**
2	**2**	**1.76**	0.45	0.84	0.81	**26-Flexibility**	-0.77	**9-Hydro(β)**
	1	0.116	3.83	2.31	0.7	10-Volume	0.68	16-Inner beta sheet
3	**3**	**2.07**	0.25	0.98	0.81	**26-Flexibility**	-0.77	**9-Hydro(β)**
	2	0.351	2.18	2.31	0.77	10-Volume	-0.63	5-Beta turn freq.
	1	0.0255	20.85	4.18	-0.71	**24-Polarity**	0.60	1-Extended Str.
4	**4**	**2.37**	0.14	7.31	0.83	**26-Flexibility**	-0.78	**9-Hydro(β)**
	3	0.649	0.85	0.01	-0.63	22-Principal z3	-0.63	3-Free E of soln.
	2	0.209	4.26	2.67	-0.54	4-Chg. Transfer	-0.52	14-Heat capacity
	1	0.00992	12.36	3.25	-0.60	**24-Polarity**	-0.48	6-C term helix
5	**5**	**2.74**	0.19	10.15	0.81	**26-Flexibility**	-0.77	**9-Hydro(β)**
	4	0.914	1	1.04	0.75	27-Flexibility	0.71	17-Hydrophobicity
	3	0.324	3.51	2.45	-0.73	15-S bend freq.	0.61	10-Volume
	2	0.0789	9.36	2.86	-0.53	**24-Polarity**	0.53	11-Beta sheet freq.
	1	0.00611	20.28	4.22	-0.64	**24-Polarity**	-0.48	6-C term helix
6	**6**	**3.67**	1.25	2.45	-0.73	2-Polarizability	0.71	**26-Flexibility**
	5	1.24	0.54	0.04	0.72	**26-Flexibility**	-0.70	**8-Hydro(α)**
	4	0.523	3.46	2.25	-0.69	3-Free E of soln.	-0.67	22-Principal z3
	3	0.192	5.13	2.93	0.56	16-Inner beta sheet	0.50	12-Width of side chain
	2	0.0474	14.68	3.37	-0.65	**24-Polarity**	-0.62	21-Hydropathy loss
	1	0.00368	21.68	4.71	-0.67	13-Positive charge	-0.51	6-C term helix
7	**7**	**3.65**	1.25	2.35	-0.73	2-Polarizability	-0.72	4-Chg. Transfer
	6	1.44	0.41	0.05	0.78	**26-Flexibility**	-0.77	**9-Hydro(β)**
	5	0.714	2.97	2.56	-0.77	3-Free E of soln.	-0.69	22-Principal z3
	4	0.333	3.61	2.37	-0.72	15-S bend freq.	0.65	12-Width of side chain
	3	0.13	12.36	3.75	-0.54	4-Chg. Transfer	0.47	18-Beta sheet freq.
	2	0.0336	18.76	3.61	-0.63	**24-Polarity**	0.59	11-Beta sheet freq.
	1	0.00277	11.75	3.05	-0.61	13-Positive charge	-0.45	6-C term helix
8	**8**	**4.51**	1.81	2.51	-0.76	4-Chg. Transfer	-0.71	20-Accessible area
	7	1.75	0.46	0.09	-0.81	**9-Hydro(β)**	0.78	**26-Flexibility**
	6	0.915	0.66	0.08	-0.63	3-Free E of soln.	-0.54	22-Principal z3
	5	0.478	8.57	3.24	-0.60	3-Free E of soln.	0.56	25-Isoelectric point
	4	0.228	6.12	3.36	-0.66	7-C term non beta	0.62	23-Hydrophobicity
	3	0.0898	13.02	3.48	-0.72	**24-Polarity**	-0.67	19-Hydrophobicity
	2	0.0233	15.89	3.57	0.58	1-Extended Str.	-0.58	24-Polarity
	1	0.00193	22.39	4.58	-0.57	6-C term helix	-0.49	24-Polarity

Table 1: Parameters for the various site classes. λ reflects the relative importance of conserving a given property (large. λ) vs. improving that quantity (small λ). Also shown are the two physico-chemical properties with the largest absolute value of correlation coefficients (cc) with $F(A_i)$. The most important of these correspond to flexibility, polarity, hydrophobicity for beta proteins (Hydro (β)), hydrophobocity for alpha proteins (Hydro (α)). The full citation of these and all other properties are given in Table 2.

The most important of the site class parameters are the fitness values. These values show which amino acids are favored in a given site class. In order to interpret this information we searched for correlation between fitness values and amino acid properties. Table 1 shows these correlation coefficients for runs with different numbers of site classes. Looking at these values, we see two main site classes with high correlation to certain properties regardless of the number of site classes. These properties are flexibility and hydrophobicity in one case and polarity in the other. Interpreting these results together with the color strips we see that the fitness values for the site class that holds the non-transmembrane residues show a strong positive correlation to flexibility and negative correlation to hydrophobicity. The fitness values of the site class that is mainly occupied by residues from transmembrane regions show a negative correlation to polarity. These correlations are in agreement with the general expectation of non-transmembrane residues being hydrophilic and transmembrane residues being hydrophobic. The positive correlation with high flexibility also makes sense since the non-membrane regions have to be highly flexible in order to accommodate the movements of the helical regions during activation of the receptor.

Code	AAindex Code	Property
1	BURA740102	Normalized frequency of extended structure
2	CHAM820101	Polarizability parameter
3	CHAM820102	Free energy of solution in water, kcal/mole
4	CHAM830108	A parameter of charge transfer donor capability
5	CHOP780101	Normalized frequency of beta-turn
6	CHOP780205	Normalized frequency of C-terminal helix
7	CHOP780211	Normalized frequency of C-terminal non beta region
8	CIDH920101	Normalized hydrophobicity scales for alpha-proteins
9	CIDH920102	Normalized hydrophobicity scales for beta-proteins
10	COHE430101	Partial specific volume
11	CRAJ730102	Normalized frequency of beta-sheet
12	FAUJ880105	STERIMOL minimum width of the side chain
13	FAUJ880111	Positive charge
14	HUTJ700101	Heat capacity
15	ISOY800105	Normalized relative frequency of bend S
16	KANM800104	Average relative probability of inner beta-sheet
17	LEVM760101	Hydrophobic parameter
18	PALJ810112	Normalized frequency of beta-sheet in alpha/beta class
19	PRAM900101	Hydrophobicity
20	RADA880106	Accessible surface area
21	ROSM880103	Loss of Side chain hydropathy by helix formation
22	WOLS870103	Principal property value z3
23	ZIMJ680101	Hydrophobicity
24	ZIMJ680103	Polarity
25	ZIMJ680104	Isoelectric point
26	VINM940102	Normalized flexibility parameters (B-values) for each residue surrounded by none rigid neighbours
27	VINM940103	Normalized flexibility parameters (B-values) for each residue surrounded by one rigid neighbour

Table 2. Amino acid indices from AAindex database[20].

5 Discussion

The preliminary results presented in this paper show that our model is capable of detecting general structural and functional differences among different locations of a protein from sequence data. The key points in this approach are the degree of relation among the proteins of interest and the phylogenetic tree that relates them. There is much evidence for a strong probability of kinship among all GPCRs. They are believed to share a common ancestor, which is believed to have given rise to general topology of seven transmembrane helices upon gene duplication[21]. Given this evidence of evolutionary relation among GPCRs and the strength of maximum likelihood methods in phylogenetics, we believe that the tree used in this study can be considered “reasonable”, if not the most likely tree. We believe that this is a good enough tree to optimize the parameters of the model, given the other studies showing low variation among estimated parameters of a model using a set of possible phylogenetic trees[22].

Besides the structural information, the model used also gives insight about the process of evolution. For all of the site classes the optimized parameters of the gaumoidal fitness function weight it towards a Gaussian distribution. This might indicate that evolution favors mutations resulting in small fitness changes, as would be expected if multiple substitutions in nearby residues favor the current amino acid type.

The most important point is that we impose no structural or functional information into the model *a priori*. All results regarding fitness and posterior values are a result of the optimization process and their correlation to structural features is a validation of the model. There is still much to do using larger data sets and site class numbers and as the resolution increases we should be able to pick more detailed structural and functional information from such studies. Currently we are only able to detect correlation and classification of general structural features such as transmembrane/non-transmembrane, but as we develop the method further we hope to be able to detect distinct site classes, holding functionally important residues. It will be interesting to compare the results of runs from other subfamilies of GPCRs and see whether these can account for their different functional properties such as coupling to different G-proteins.

Acknowledgments

Thanks to Sarah E. Ingalls for her contributions in writing the software and Todd Raeker for computer support. The financial support was provided by NIH Grant LM0577, NSF Grant 9726427, and NSF equipment grant BIR9512955.

References

1. Iismaa PT, B.J., Shine J, *G-Protein Coupled Receptors*. 1995: Springer Verlag.
2. Palcewski K, K.T., Hori T, Behnke CA, Motoshima H, Fox BA, Trong IL, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M, *Crystal Structure of Rhodopsin: A G-protein Coupled Receptor.* Science, 2000. **289**: p. 739-745.
3. Attwood, T., *Acompendium of specific motifs for GPCR subtypes.* Trends in Pharmacological Sciences, 2001. **22**(4): p. 162-165.
4. Horn F, W.M., Oliviera L, Ijzerman AP, Vriend G, *Receptors Coupling to G proteins: Is There a Signal Behind Sequence?* Proteins: Structure, Function and Genetics, 2000. **41**: p. 448-459.
5. Dayhoff M. O., E.R.V., *Atlas of Protein Sequence and Structure*, . 1966, National Biomedical Research Foundation.
6. Jones D.T., T.W.R., Thornton J.M., *The rapid generation of mutation data matrices from protein sequences.* Computational Applications in Bio Sciences, 1992. **8**(3): p. 275-282.
7. Yang Z., *Maximum-Likelihood Estimation of Phylogeny from DNA Sequences When Substitution Rates Differ over Sites.* Molecular Biology and Evolution, 1993. **10**(6): p. 1396-1401.
8. Koshi JM, G.R., *Context Dependent Optimal Substitution Matrices.* Protein Engineering, 1995. **8**: p. 641-645.
9. Goldman N., T.J., Jones DT., *Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution.* Genetics, 1998. **149**: p. 445-458.
10. Lio P, G.N., *Using Protein Structural Information in Evolutionary Inference: Transmembrane Proteins.* Mol. Biology and Evolution, 1999. **16**(12): p. 1696-1710.
11. Halpern AL, B.W., *Evolutionary Distances for Protein-Coding Sequences: Modeling Site-Specific Frequencies.* Mol. Biology and Evolution, 1998. **15**(7): p. 910-917.
12. Yang Z, *Relating Physicochemical Properties of Amino Acids to Variable Nucleotide Substitution Patterns Among Sites.* Pacific Symposium on Biocomputing, 2000.
13. Yang Z., N.R., Goldman N., Pedersen AM., *Codon-substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites.* Genetics, 2000. **155**: p. 431-449.
14. Koshi JM, M.D., Goldstein RA, *Using Physical-Chemistry-Based Substitution Models in Phylogenetic Analyses of HIV-1 Subtypes.* Mol. Biology and Evolution, 1999. **16**(2): p. 173-179.

15. Dimmic MW, M.D., Goldstein RA, *Modeling Evolution at the Protein Level Using an Adjustable Amino Acid Fitness Model.* Pacific Symposium on Biocomputing, 2000: p. 18-29.
16. Yang Z., *Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods.* Journal of Molecular Evolution, 1994. **39**(306-314).
17. Horn F, W.J., Beukers MW, Horsch S, Bairoch A, Chen W, Edvardsen O, Campagne F, Vriend G, *GPCRDB: an information system for G protein coupled receptors.* Nucleic Acid Research, 1998. **26**(1): p. 275-279.
18. Adachi J., H.M., *Maximum Likelihood Inference of Protein Phylogeny*, . 1992: Tokyo.
19. Yang Z., *Phylogenetic Analysis by Maximum Likelihood*, . 2000, University College, London: London.
20. Kawashima S., O.H., Kanehisa M., *AAindex: amino acid index database.* Nucleic Acid Res., 1999. **27**: p. 368-369.
21. Taylor EW, A.A., *Sequence Homology Between Bacteriarhodopsin and G-protein Coupled Receptors: Exon Shuffling or Evolution by Duplication.* FEBS, 1993. **325**(3): p. 161-166.
22. Yang Z., *A Space-Time Process Model for the Evolution of DNA Sequences.* Genetics, 1995. **139**: p. 993-1005.