Pacific Symposium on Biocomputing 13:426-437(2008)

COMPARISONS OF PROTEIN FAMILY DYNAMICS

A. J. RADER* and JOSHUA T. HARRELL

Department of Physics, Indiana University Purdue University Indianapolis,
402 N. Blackford St., LD156D
Indianapolis, IN 46219, USA
**E-mail: ajrader@iupui.edu*
www.physics.iupui.edu/∼ajrader/

Similarities between different protein structures have led to the identification of protein families based upon some measure of structural similarity. Using these similarities one can classify proteins into structural families and higher-order groupings from which inferred function can be transferred. When taken for a large number of proteins, these schemes point to evolutionary relationships between organisms. We propose a novel classification scheme based upon the structurally-inspired dynamics of each protein. This classification scheme has the advantages of being quantitative, automatically assigned, and able to also make distinctions within protein families. Results are presented for five protein families illustrating the correct identification of previously un-classified structures and sources of intrafamily distinctions.

Keywords: GNM; protein dynamics; conformations; families.

1. Introduction

The comparison of proteins from different organisms relies heavily upon the paradigm that sequence encodes for protein structure which in turn determines protein function. Often protein function is not a easily definable quantity[1] making some associations unreliable. More directly, proteins can be grouped into families based upon shared structural characteristics since structural changes are generally more conservative than sequence changes. Two widely used structurally-based classification systems are SCOP (Structural Classification of Proteins)[2] and CATH (Class, Architecture, Topology, and Homologous superfamily).[3] Both classification systems require some manual intervention and depend upon the additional step of defining the domains within a protein structure. The assignment of such domains is not a unique process and adds another layer of complication to such classifications.

We contend that structurally-inspired information, specifically protein *dynamics*, are important for making the correct *functional* assignment of proteins.[4] Information regarding dynamics is absent in both of the two structural classification systems mentioned above. In this paper we present an automatic assignment criteria for grouping protein families based upon their entire structure rather than the added step of domain identification. Thus although the analysis presented here is similar to the SCOP and CATH classifications it differs by considering the dynamics of complete protein structures.

The Gaussian Network Model (GNM)[5,6] provies an efficient calculation of protein dynamics by representing the protein structure by an elastic network of residues. This creates a coarse-grained representation of the structure and its dynamics. As a result, comparison of low frequency (global) modes of motion from GNM to proteins with a similar Rossmann-fold displayed a striking similarity.[7] A related study applied to the globin family observed a similar trend that similar protein structures exhibited similar dynamics.[8] Preliminary analysis at the superfamily level found that regions with high mobility also demonstrated high levels of evolutionary fluctuations.[9] In this work we quantify the degree of similarity in dynamics with the aim of exploring how these dynamics play a role in defining protein function. We generalize these comparisons to families of proteins based upon the SCOP classification schemes, allowing a new automatic classification of each protein in terms of their GNM-defined dynamic similarities.

2. Methods

2.1. *Protein Family Selection*

The Protein Data Bank (PDB)[10] was used in conjunction with the iGNM database[11] to select the families of proteins used in this study. The iGNM database is an online resource of pre-computed GNM analysis for all structures deposited in the PDB. The low frequency eigenvectors, termed slow modes, were used in the analysis presented here because these slow modes have previously been associated with global motions and likely (large-scale) functional motions.[12]

The SCOP (v 1.71) classification was used as the inital basis for familial groupings. Five families were chosen from the SCOP database site such that the proteins in these families represented different functions, architectures and number of residues. A family was considered if it had more than 25 member structures with the same number of residues. Since the number of structurally resolved residues does not always match the protein sequence length, the list of SCOP family member PDB structures was checked against the iGNM database to determine the number of nodes (residues) present in each structure. Only structures with the same number of residues were selected for use in this study. Requiring the proteins to have the same length allowed a direct comparison of them against each other using the dot product of their modes of motion (see below). Once this number of residues was determined, an additional set of structures was obtained by retrieving all structures present in the iGNM with this number of residues. Thus each protein family studied had a set of structures already deemed part of the (SCOP) family and a second set of (non-family) structures each having the same length as those in the family. The analysis was carried out for the five SCOP families listed in Table 1.

Table 1. List of protein families tested.

Abbreviation: Name	Residue count	Family	Non-family
FABP: Fatty acid binding protein-like	131	30	42
Glob: Globins	153	84	58
CytC: monodomain cytochrome c	108	28	70
DHFR: Dihydrofolate reductases	159	27	46
PoBP: Phosphate binding protein-like	517	30	6

2.2. *GNM*

The GNM[5,6] treats the structure as an elastic network model where amino acid residues within a cutoff distance, r_c are connected by springs with a uniform force constant. In this model, the C^α atom positions of each residue serve as the nodes. Denoting R_{ij} as the distance between residues i and j, a Kirchhoff or connectivity matrix, $\mathbf{\Gamma}$, is constructed such that off-diagonal elements are -1 when $R_{ij} \leq r_c$ and 0 when $R_{ij} > r_c$; while the diagonal elements are the sum of off-diagonal elements. The normal modes characterizing the motion of this network are found by eigenvalue decomposition of the Kirchhoff matrix according to Eq. (1)

$$\mathbf{\Gamma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{1}$$

where $\mathbf{U}$ is a matrix composed of eigenvectors, $\mathbf{u}_i$ $(1 \leq i \leq N)$, and Λ is the diagonal matrix of the eigenvalues λ_i. Despite being a purely topological model, GNM and related models have been widely used to characterize functionally relevant motions in terms of a few low frequency (small λ) modes.[4]

For each of the structures in the protein families, the 20 slowest mode (lowest frequency) eigenvectors were downloaded from the iGNM database. The correlation between residue fluctuations ($\Delta\mathbf{R}_i$) due to a specific mode is calculated according to Eq. (2).

$$[\Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j]_k = \frac{3k_bT}{\gamma}\lambda_k^{-1}[\mathbf{u}_k]_i[\mathbf{u}_k]_j \tag{2}$$

Here $[\mathbf{u}_k]_i$ is the i^{th} element of the eigenvector $\mathbf{u}_k$, λ_k is the eigenvalue, T is the absolute temperature and k_b is the Boltzmann constant.

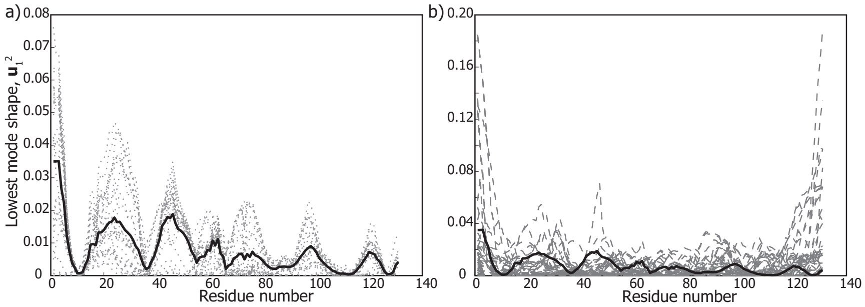

Fig. 1. Relationships between mode shapes. a) The lowest mode shape $[\mathbf{u}_1]^2$ is plotted for each family structure with gray dashed lines. The thick black line highlights the average mode shape. b) The lowest mode shape for each non-family structure is plotted in gray dashed lines and contrasted with the average family mode shape in black.

Figure 1 illustrates the lowest mode shape $[\mathbf{u}_1]^2$ plotted against residue number. By inverting Eq. (2) one can see that this is proportional to the lowest mode self-fluctuation or mobility. In Fig. 1a) the results are plotted for each of the *family* protein structures as gray dashed lines. One can see that each structure has a slightly different degree of mobility in this plot. Clearly there are some regions of qualitative agreement such as the coincidence of minima, highlighted by the average curve in black. It has previously been demonstrated that these minima serve as *hinge* sites that correlate with binding and/or catalytic sites.[12] Complimenting this previous insight, we observe that the largest variations occur not in these hinge sites but in the mobile regions between such hinges. As shown here, the general mode shape illustrates the ability of GNM to cluster groups of structures by their dynamics. Additionally the variations in the degree of modal mobility points to the ability of GNM to differentiate between similar structures. Figure 1b) shows the same average slow mode mobility in black compared results for each of the non-family structures (dashed lines). Unlike the case of structures from a family, there is no observable trend for mobility among non-family structures. Beyond this qualitative observation, we desire to develop a more quantitative comparison for a large number of structures.

2.3. *Quantitative Mode Comparisons*

Letting $\mathbf{x}_\alpha$ and $\mathbf{y}_\beta$ represent the α^{th} and β^{th} eigenvectors of proteins x and y respectively, one can define the dot product between eigenvectors of different proteins according to Eq. (3).

$$P_{\alpha\beta}^{xy} = \mathbf{x}_\alpha \cdot \mathbf{y}_\beta = \sum_{i=1}^{N} x_{\alpha i} y_{\beta i} \tag{3}$$

Using the fact that these are eigenvectors, we ignore elements of Eq. (3) corresponding to the same protein. Thus if k represents the number of eigenvectors being considered we define a $(k \times k)$ matrix for each pair of proteins in the dataset given by Eq. (4).

$${}_k\mathbf{P}(x,y) = \begin{bmatrix} P_{11}^{xy} & P_{12}^{xy} & \cdots & P_{1k}^{xy} \\ P_{21}^{xy} & P_{22}^{xy} & \cdots & P_{2k}^{xy} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k1}^{xy} & P_{k2}^{xy} & \cdots & P_{kk}^{xy} \end{bmatrix} \tag{4}$$

Thus when we compare L proteins, we can define a large $(kL \times kL)$ matrix comprised of smaller $(k \times k)$ matrices, which compare the individual k lowest modes of each protein against the k slowest modes of the other proteins in the family set. The amount of correlation data contained here makes it hard to recognize the correlations between proteins. In order to succintly compare the data, two correlation metrics are introduced as functions of the number of eigenvectors being compared, denoted by k. The first correlation, $M_k(x,y)$, defines the maximum dot product between the slowest k modes for each pair of structures (Eq. (5)) and the second (Eq. (6)) calculates the sum of the maximum values for each column, j, in the smaller $(k \times k)$ matrix.

$$M_k(x,y) = \max\left(|{}_k\mathbf{P}(x,y)|\right) \tag{5}$$

$$S_k(x,y) = \sum_{j=1}^{k} M_j(x,y) \tag{6}$$

Two different measures of correlation were introduced because it is not clear how correlations between different modes in different structures should be computed *a priori*. The M_k correlation measure is concerned with identifying *any* two highly correlated eigenvectors between the two proteins without regard for the specific order of these low frequency modes. This acccounts for the possibility of mode *mixing* which occurs when the lowest eigenvector from one protein is highly correlated with an eigenvector from another protein that is not the lowest frequency mode. In contrast, the second correlation measure, S_k, focuses on identifying how well the *entire* subspace spanned by the low frequency eigenvectors of one protein matches the subspace spanned by the low frequency eigenvectors of another protein.

By averaging these measures over the set of family structures we can determine the average amount of correlation a structure has with respect to a protein family. For a family with l_f proteins, the family averaged, M_k value for the x^{th} protein, $\langle M_k(x)\rangle_f$, is defined by Eq. (7) along with a family averaged standard deviation, $\langle \sigma_k^M \rangle_f$.

$$\langle M_k(x)\rangle_f = \frac{1}{l_f}\sum_{y=1}^{l_f} M_k(x,y) \tag{7}$$

Similarly, the family averaged, S_k value for the x^{th} protein, $\langle S_k(x)\rangle_f$, is defined in Eq. (8) along with an average standard deviation, $\langle \sigma_k^M \rangle_f$.

$$\langle S_k(x)\rangle_f = \frac{1}{l_f}\sum_{y=1}^{l_f} S_k(x,y) \tag{8}$$

3. Results

3.1. *Classification of Protein Families by Dynamics*

In order to determine a suitable number of modes to consider we performed calculations for all values of $k \leq 20$. Taking the average of the family averaged metrics in Eq. (7) and Eq. (8), defines an overall correlation value for each protein family.

$$\langle\langle M_k\rangle_f\rangle = \frac{1}{l_f}\sum_{x=1}^{l_f} \langle M_k(x)\rangle_f \tag{9}$$

Figure 2 plots these family-averaged values ($\langle\langle M_k\rangle_f\rangle$ and $\langle\langle S_k\rangle_f\rangle$) as a function of the number of modes, k. The M_k averages when $k = 1$ are relatively low (0.67) in the case of FABP. This is due to the fact that using only one eigenvector from each protein prevents considering the correlation

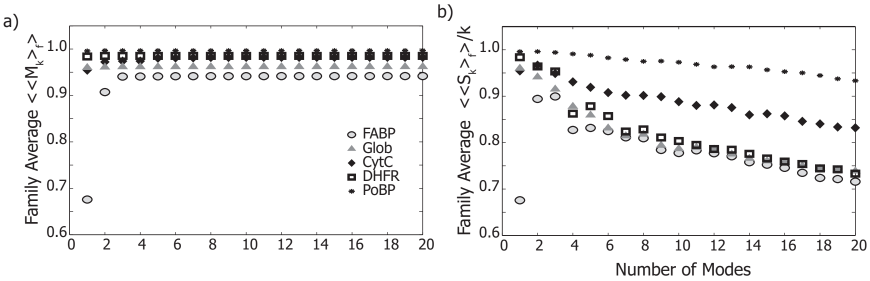

Fig. 2. The family averaged metrics calculated for different numbers of modes. a) $\langle\langle M_k\rangle_f\rangle$ b) $\langle\langle S_k\rangle_f\rangle/k$

between potentially mixed modes. However, as more modes are included, this situation is quickly remedied. By the time $k = 5$, the value of $\langle\langle M_k\rangle\rangle$ has reached an asymptote and so $k = 5$ was chosen for the results presented here. Due to the high average family correlation expressed by M_5, this measure suggests a means to distinguish family from non-family structures.

The trend for $\langle\langle S_k\rangle_f\rangle/k$ in Fig. 2b) is different, generally reflecting the a greater disparity for larger values of k. However, as illustrated by FABP, more than one mode is required to account for potential mode mixing. The overall lower correlation of S_k when compared to M_k makes S_k more appropriate for monitoring distinctions *within* a family. To keep the analysis consistent with M_k and allow for intrafamily distinctions, we set $k = 5$ for the S_k results presented here.

We plot the $M_k(x, y)$ correlation values between each pair of protein structures using a scheme that runs from no (zero) correlation in blue to maximal (one) correlation in red. Figure 3 shows this plotted for FABP, Glob and CytC in panels a) through c) respectively with each row and column corresponding to a specific protein structure. The proteins identified by SCOP as family members are plotted first in each case. Similarly, we plot the $S_k(x, y)$ values with a color scheme ranging from no correlation in blue to maximal correlation in red. Figure 3d) through f) plots this correlation for the family members of these three families. Results for the other two protein families (DHFR and PoBP) are similar and thus not shown explicitly. As mentioned above, these results are shown for $k = 5$ although the calculations were repeated for all values of $k \leq 20$.

To understand the significance of these plots, consider CytC (Fig. 3c) as an example. The first 28 rows and columns correspond to the 28 CytC family proteins and are shown in shades of red to signify their high corre-

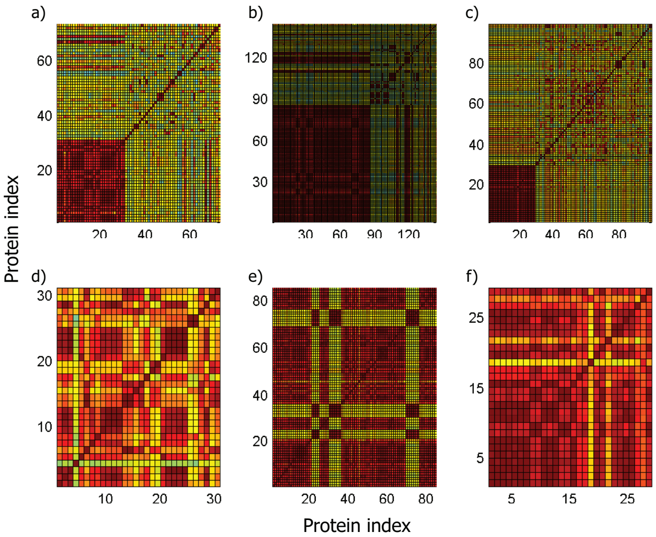

Fig. 3. Correlation measure ($k = 5$) plots for three protein families. a) M_5 for FABP b) M_5 for Glob c) M_5 for CytC. The known family members are listed first followed by other proteins outside the family. In each case there are clear distinctions between family and non-family structures. High correlation corresponds to red and low correlation is in blue. For the S_5 plots, only the subset corresponding to family members is plotted to show the intra-family distinctions. d) S_5 for FABP e) S_5 for Glob f) S_5 for CytC.

lation. By contrast, the last 70, non-family structures (rows and columns) are in shades of greens and yellows indicating low correlation (0.3 to 0.6). This distinction in color clearly shows a difference between the modes of proteins within the SCOP family compared to the modes of proteins not in the family. This trend is observed for each of the five protein families studied. Such a trend suggests M_k correlations of protein dynamics can be used as a classifying technique.

The clear distinction between M_k values for proteins in the family versus non-family proteins makes idenfication of potential *candidates* for inclusion in the family relatively simple. There are some proteins in Fig. 3a) and b) not classified as being part of the SCOP family that share the same color pattern as those within the family. This indicates a high degree of correlation between the eigenvectors from these structures and those in the family implying that they should be considered as part of the family.

3.2. *Identification of Dynamically Similar Proteins*

The numerical values behind the similar color patterns are used to identify these candidate structures with respect to the family average $\langle M_k(x)\rangle_f$ values defined in Eq. (7). Specifically, we consider a structure as a family *candidate* if its family averaged value is within three standard deviations of the mean family averaged value as defined in Eq. (10).

$$\langle M_k(x)\rangle_f \geq \langle\langle M_k\rangle_f\rangle - 3\langle\sigma_k^M\rangle_f \qquad (10)$$

Thus instead of looking at the correlation color patterns illustrated in Fig. 3, we can plot the family averaged, $\langle M_k(x)\rangle_f$, values against protein index as in Fig. 4a). Indicatting the $3\sigma_5^M$ limits by gray dashed lines quickly identifies three potential candidates for the FABP family according to the criteria in Eq. (10). These three candidate structures have protein indices 57, 66 and 68 corresponding to PDBids 1t8v, 1yiv and 2a0a.

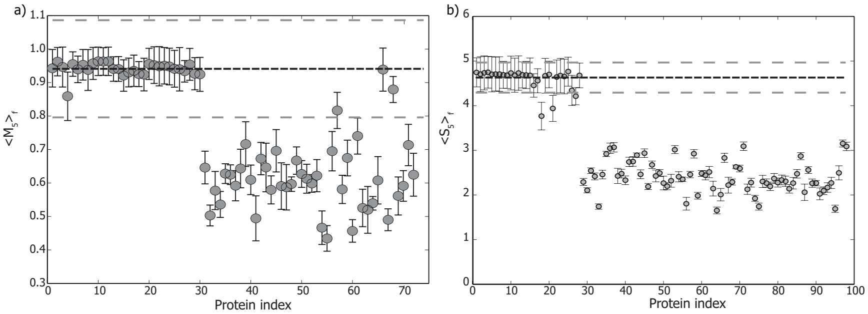

Fig. 4. The family averaged correlation measures for different proteins in a family illustrate the candidate and outlier criteria. (a) $\langle M_5(x)\rangle_f$ values for each FABP family and non-family structure. The mean family value, $\langle\langle M_5\rangle_f\rangle$, is shown by a black dashed line and the candidate $3\sigma_5^M$ range is indicated by gray dashed lines. (b) $\langle S_5(x)\rangle_f$ values for each CytC family and non-family structure. The mean family value, $\langle\langle S_5\rangle_f\rangle$, is shown by a black dashed line and the outlier σ_5^S range is indicated by gray dashed lines.

Table 2 summarizes the mean family averaged values as well as the number of candidates for each family. In the case of FABP, none of these three structures are part of the SCOP family. 1t8v is a fairly recent PDB structure which is annotated as a fatty acid binding protein, but due to its recent deposition it has not been included in the SCOP database. The other two structures, 1yiv and 2a0a, are annotated as a myelin protein and dust mite allergen respectively. Visual inspection of these structures confirms that they contain the dominant β-barrel structure of FABP structures and

Table 2. Family averaged correlation metrics and standard deviations

Family	$\langle\langle M_5\rangle_f\rangle$	$\langle\sigma_5^M\rangle_f$	Candidates	$\langle\langle S_5\rangle_f\rangle$	$\langle\sigma_5^S\rangle_f$	Outliers
FABP	0.9411	0.0484	3	4.1579	0.4818	2
Glob	0.9617	0.0377	10	4.2958	0.5221	17
CytC	0.9804	0.0267	0	4.5958	0.3373	3
DHFR	0.9852	0.0104	0	4.3925	0.4174	0
PoBP	0.9967	0.0083	0	4.4045	0.0710	0

suggests a new potential functional mechanism for these structures, namely as a fatty acid binding protein.

Analysis of the Glob family indicated ten candiate structures all of which are recent myoglobin structures which are not included in the SCOP database. Correct identification of these structures as part of the Glob family by comparison of their dynamic modes serves to confirm the applicability of this method to distinguish protein families. The other three families: CytC, DHFR and PoBP do not have any candidate proteins.

3.3. *Intrafamily Distinctions*

After demonstrating the ability of this method to distinguish family from non-family structures, the ability to distinguish variations among structures within a family was also investigated. As can be seen in Fig. 3d) through f), correlations among structures within a family are not uniform but have variations. Regions with lower correlation (greens, yellows and oranges in these panels) correspond to structures that are potential family *outliers*. Similar to the definition of candidates in Eq. (10), we define such outliers as being more than one standard deviation ($\langle\sigma_k^S\rangle_f$) away from the mean family averaged S_k value as in Eq. (11).

$$\langle S_k(x)\rangle_f \leq \langle\langle S_k\rangle_f\rangle - \langle\sigma_k^S\rangle_f \tag{11}$$

Again the actual values of $\langle\langle S_5^S\rangle_f\rangle$ and $\langle\sigma_5^S\rangle_f$ used for each of the families are listed in Table 2. Using this outlier criteria we are able to pick out structures that may be structurally and/or functionally distinct from within a family in an automated fashion.

Beginning with CytC as an example case, one can see two green-yellow-orange bands in Fig. 3c) representing protein indices 18 and 21 that are less correlated with other CytC structures in general. Figure 4b) plots the family averaged correlation measures, $\langle S_5(x)\rangle_f$, for each protein along with the outlier criteria from Table 2. In this plot one can see that all non-family members fall below the $\langle\sigma_5^S\rangle_f$ level shown with a dashed gray line because there were no candidates this family. More importantly, there are three outliers corresponding to protein indices 18,21 and 27. These indices refer to PDBids 1fhb, 1nmi and 1yic. Since these *outliers* have the same overall family structure, the differences identified by this analysis of the dynamics are due to some other factor(s). Using similar analysis, FABP had two outliers: PDBids 1ael and 1tou (indices 4 and 25 in Fig. 3a) and Glob had 17 outliers identified (protein indices 21–24, 30–35 and 69–75 in Fig. 3b).

Examining the structures that were deemed to be outliers as a whole we are able to determine a few reasonable explanations for these differences. The outliers for FABP and CytC were structures that were determined by NMR rather than X-ray crystallography. Although these were not the only NMR structures in the datasets representing these families, it suggests that structures that are not forced to conform to the Ramachandran phi-psi plot may adopt a "looser" structure with measurable differences in dynamics. Further supporting this claim is the fact that one of the CytC outliers, 1nmi, is an averaged NMR structure which would not necessarily reflect the true dyanamics of the CytC family. The 17 outliers in the Glob family correspond to *all* of the structures of one type of globin: leghemoglobin from a specific species: yellow lupin. In this case, these structures serve to form a sub-family within Glob that can be seen visually by the green-yellow bands in Fig. 3b).DHFR and PoBP had no outliers according to the criteria of Eq. (11). However examination of the most varied structures support the claims of sub-family organization by speciation and differences due to ligand-binding state (data not shown).

4. Conclusions

We have demonstrated an automatic family classification scheme for protein structures based upon their computed dynamics. Comparisions using the low frequency eigenvectors of structures accurately assigns these structures to a unique protein family. Using this precomputed data, Eq. (10) provides a measure for assigning newly determined structures as candidates to a particular protein family.

In addition, this method provides a quantitative measure of the differences within protein families. These differences can be investigated in terms of outliers or most dynamically different as indicated in the text. Examination of the outliers indicates that differences within the familes can be attributed to some combination of differences in ligand-binding state, method of structural determination and sequence. These factors are an initial list of