An Introduction To Motif Based Functional Classification of Large Protein Families

Many methods of clustering proteins within large protein families either build up from pairwise sequence alignments or rely solely on hierarchical clustering methods. While these methods can be incredible useful, they may not efficiently discover small regions of similarity in large multidomain proteins, and they may miss functional similarities that arose due to domain shuffling or convergent evolution rather than due to divergent evolutionary processes. Here, we describe the rationale behind motif based functional classification, and the benefits of including non-hierarchical methods in such analyses. We provide a general review of one new program, CASTOR, which uses a motif based approach to elucidate both hierarchical and nonhierarchical relationships within protein families. The application of this method to classification of the large mammalian G protein coupled receptor family, and to the family of approximately 1000 known mouse and human olfactory receptors, has demonstrated the utility of such programs in predicting functional relationships that may be missed by standard multiple sequence alignments or phylogenetic analyses. Finally, we suggest the application of motif based functional classification to a family of relatively uncharacterized proteins, the vomeronasal receptors.

MOTIF BASED FUNCTIONAL CLASSIFICATION OF LARGE PROTEIN FAMILIES

The assignment of functions to uncharacterized proteins is a complicated and challenging task. As a first step in assigning function, the sequence of an uncharacterized protein can be compared to a database of proteins with known functions, and sequence similarities can be used to assign new proteins to existing protein families. The presence of similar domains implies a similar biochemical function or structure, and a group of similar sequences may define a family of proteins that may share a common evolutionary or biochemical origin (Mount 2001).

One major challenge is to develop effective methods for assigning specific functions to proteins within large families and for classifying family members into groups that share functional characteristics. While multiple sequence alignments can be incredibly useful in predicting functionally or structurally important regions that are shared by a family of protein sequences (Gotoh 1996, Heger and Holm 2000), problems may arise when more than one domain is present in a protein (Mulder and Apweiler, 2001). In this case, it is possible that similarities found to one domain of a given protein may end up masking or overwhelming similarities to another domain. This problem is especially interesting in the case of families of multidomain proteins, such as the G protein coupled receptors (GPCRs). To illustrate some of the difficulties involved, consider the simplified example of a hypothetical family of multidomain proteins in which each family member has 14 domains. Seven of these domains are very large and encode transmembrane regions, two smaller domains encode regions involved in downstream signaling, and the remaining five domains encode small regions involved in forming the ligand binding pocket. If these proteins are characterized solely based on sequence identity or sequence similarity, it is possible that functional similarity between two sequences that share regions involved in ligand binding but have divergent sequences in the longer transmembrane domains may be missed. Conversely, proteins with little or no functional similarity may be predicted to be similar based on the presence of extensive sequence similarity in regions that do not affect function.

An alternative method of discovering similarity of function between proteins within large multidomain families (instead of building up from pairwise sequence alignments) is to base the measure of similarity on the presence or absence of motifs that are likely to have functional relevance. While there are extensive databases of known patterns and profiles that have been built from alignments of related sequences (Mulder and Apweiler 2001), novel sets of statistically significant motifs can be generated based on a the members of a given family and then can be used in analysis of that family. Two extremes in this type of classification are the binary tree approach and the nonrecursive graph approach (Liu and Califano 2003). In the binary tree approach, a number of motifs are inferred from a family of sequences S, and are ranked according to statistical significance. The most statistically significant motif is then used as a basis for dividing *S* into a positive set (sequences that contain the motif) and a negative set (sequences that do not contain the motif). The first motif is then masked in the positive set, and the two sets are separately analyzed for the presence of additional statistically significant motifs. This is done recursively until statistically significant motifs can no longer be discovered. The result is a hierarchy of classes that can be viewed as a binary tree in which for any node (each node corresponds to a class) one motif is discovered and used to infer both of its children (the positive and negative sets with respect to that motif) (Liu and Califano 2003). If two variables N_B and N_D describe the bredth (meaning the number of statistically significant motifs that will be used to partition a given set before moving on to analyze the children of the set) and the depth (meaning the number of levels of recursion which will be allowed to occur) of the classification strategy, then the binary tree can be described by the values ($N_B = 1$, N_D = infinity). In the nonrecursive graph approach, again motifs are found in *S* and are ranked according to statistical significance. This time, however, after the first motif is selected and used to define positives and negatives, it is masked and the process is repeated without any splitting of *S*. This type of analysis of *S* is repeated as long as distinct, statistically significant motifs can be generated from *S*. The classes defined by all the negative and positive sets (each with respect to a different motif), may be partially or completely overlapping. The result is a flat list of possibly overlapping subsets. This approach can be described by the values (N_B = infinity, $N_D = 1$) (Liu and Califano 2003).

These two extreme models for motif based functional classification both have strengths and weaknesses. The binary tree, which is a hierarchical structure, represents divergent evolutionary processes very well. However, it does not do well with complex, nonhierarchical relationships among functional subsets that are the results of domain shuffling and convergent evolution (Liu and Califano 2003). For example, when the most statistically significant motif is used to divide the family into two initial branches (one that has the motif and one that does not) and further analyses focus exclusively on members of one branch or the other, then the presence of a common, less statistically significant motif in members of the two different branches will be missed. The nonrecursive graph model, on the other hand, creates a nonhierarchical structure that can represent relationships caused both by divergent evolutionary processes and by domain shuffling or convergent evolution. The main drawback of this approach is that without recursive processing, small functional subsets may be missed because the motifs that characterize them cannot reach statistical significance when the full set of sequences *S* is used as the basis for every search (Liu and Califano 2003).

CASTOR: A NEW PROGRAM THAT UTILIZES MOTIF BASED FUNCTIONAL CLASSIFICATION

Liu and Califano have recently created a program, CASTOR, which uses statistically significant motifs to identify protein regions likely to have functional significance and then classifies proteins within a family based on these regions (Liu and Califano 2003). This program differs from many similar programs in that rather than using a hierarchical bottom up clustering method based on pairwise sequence similarity, CASTOR uses motifs to infer likely protein subsets in a top-down and recursive manner. This allows discovery of both hierarchical and non-hierarchical subset relationships (Liu and Califano 2003). CASTOR has been tested on the family of G protein coupled receptors (GPCRs) and was found to generate an organization of the family that agrees incredibly well with an organization based on existing biological knowledge of the family. The details of the CASTOR algorithm are available (Liu and Califano 2003), and will be summarized here.

Overview of the CASTOR program: The input for CASTOR is a group of related sequences, *S*. In what the authors describe as a "recursive graph model," the user can set values for the number of distinct motifs (N_B) to which pattern discovery is limited for any given class, and for the number of recursion levels with respect to *S* (N_D) that will be allowed. The binary tree model discussed above is described by ($N_B = 1$, N_D = infinity), while the nonrecursive graph model is described by (N_B = infinity, $N_D = 1$). According to the recursive graph model, up to $N_B + 1$ classes (F_1 through F_{NB}, plus a class F_0 that contains none of the discovered motifs) may be inferred by pattern discovery from the full set S. This is done by first searching the set *S* for motifs, then using presence of the most statistically significant motif as a requisite for sequences in *S* to be considered members of class F_1. This first motif is then masked, and the process is repeated to define class F_2, and so on. Since sequences that are part of F_1 are included in further rounds of motif discovery, the classes may overlap substantially (meaning that a given sequence may be a member of several different classes). Once N_B classes have been defined or there are no remaining statistically significant motifs to be discovered from *S*, the same type of analysis is repeated for each of the daughter classes of *S* (F_1 through F_{NB}, plus F_0). As an example, from class F_1 an additional $N_B + 1$ daughter classes will be created, labeled ($F_{1,0}$) through ($F_{1,NB}$). This recursive process is repeated for N_D levels. For example, if $N_D = 3$, then the process will be done once for the full set *S*, once for each of the daughters of *S* (F_0 through F_{NB}), and once for each of the daughters of F_0 through F_{ND}. The parameters N_B and N_D can be varied on a case by case basis. Since any time that $N_B > 1$, classes may overlap, a class that corresponds to the intersection of overlapping classes may later be inferred twice. (For example, if classes F_1 and F_2 share several sequences, then at the next recursive level, when classes are being inferred separately from F_1 and from F_2, a class defined by all or some of the overlapping sequences may be inferred twice, once from F_1 and once from F_2). CASTOR addresses this problem by a process termed "class space pruning." An additional problem is that, especially when $N_B = 1$, the program may discover the same motif from two non-overlapping classes. (For example, if motif B is used to divide *S* into classes F_0 and F_1, but motif A is common to F_0 and F_1, motif A may be independently discovered later on in subsets of both F_0 and F_1, leading to redundancy). CASTOR addresses this problem by a process termed "structure tidying."

Summary of methods used by CASTOR: The processes by which CASTOR discovers and refines motifs, does "class space pruning," and does "structure tidying" are critical for its method of protein classification. They will be briefly summarized here.

A. Motif discovery and refinement: CASTOR begins by finding statistically significant motifs using a program called SPLASH. The motifs are first represented as regular expressions. These provide a rough and rigid representation of the underlying motifs, and may not be flexible enough to fully characterize a functional motif. Therefore, the regular expression is extended to both sides using a sliding window that is moved across the regions flanking the motif. Additional conservation in these flanking regions is analyzed by computing "amino acid entropy values" (which are based on the frequency of amino acids at each position) at each position of the sliding window. The motif is extended until this entropy value increases beyond a predefined threshold, or the end of the sequence is reached. After this extension process, the modified motifs are used as starting points for the generation of profile hidden markov models (HMMs). The HMM is a statistical model that allows for position specific scoring of substitutions, insertions and deletions. An HMM is first generated based on the initial sequences (usng the HMMER program), and is then refined in an iterative fashion. Once the first HMM is constructed based on the initial aligned matches that define the motif (H), it is run against the full set of sequences to generate a modified set of aligned matches (H_i). A new profile HMM is then build based on the sequences in Hi. This process could be repeated until $H_{i+1} = H_i$, but here the authors chose to refine the model only once (Liu and Califano 2003).

B. Class space pruning: As mentioned above, if two classes contain common sequences (which is possible when $N_B > 1$), a class contained within the intersection of the original two classes may be inferred more than once, due to separate processing of the original two classes, and thereby lead to redundancy. This problem is addressed by comparing each class F_i (except for F_0) to see if it is a subset or a duplicate of another class F_j that was also inferred from F, and whether it is a subset or duplicate of another class $F^{=}$ that was inferred from a different class. If F_i is a subset of F_j or $F^{=}$, then F_i is discarded and the corresponding motif is discarded and unmasked, allowing it to be rediscovered later from F_j or $F^{=}$. If $F_i = F_j$, then the one inferred first is discarded, but the corresponding motif is retained, creating a single class that is characterized by two different motifs. If $F_i = F^{=}$, then the one inferred earlier is discarded (Liu and Califano 2003).

C. Structure tidying: Using the terminology introduced above, a set of sequences F is the support set of a motif M that was discovered from a set of sequences F+. As mentioned earlier, however, if $N_B <$ infinity, there may be sequences in the full sequence set *S* that do not belong to F, but which do carry motif M. This is the type of situation in which non-hierarchical relationships can be missed. To avoid this problem and to identify non-hierarchical relationships among functional subsets, CASTOR attempts to "tidy" the classes such that the classes are the full support sets of the corresponding motifs (meaning that all sequences containing a motif M are part of the class that is the support set of M). The full support set is found by running the profile HMM that represents a given motif (M) against the complete database *S*. The profile HMM is then refined once (see above), and the set of sequences that match the refined profile HMM are considered to be the full support set of motif M. All of the motifs (M_i, M_j, etc) are then compared to each other by considering how much overlap exists between multiple sequence

alignments (MSA) of their respective support sets, and motifs with substantial overlap of their MSAs are grouped together. The parameters for these comparisons may be adjusted by the user. A consensus motif, in the form of a profile HMM, is constructed for each of the motif groups and is refined once. The consensus motifs are run against *S*, and if the support sets for any of the consensus motifs are almost identical, they are merged into a class group. For example, if the support sets for consensus motifs M_x and M_z are found to be almost identical, they will be merged to form a class group in which the great majority of the member sequences will carry both M_x and M_z. A consensus class is then created for each class group (Liu and Califano 2003).

Success of CASTOR in representing complex non-hierarchical relationships among functional subsets within the GPCR family: The authors show that by varying the values of N_B and N_D, they can reveal otherwise hidden functional similarities between members of the GPCR family. For example, when the values are set to ($N_B = 2$, N_D = infinity) and the iterative motif refinement (discussed above) is used, a specific motif T is found to be supported unambiguously and exactly by the subset of GPCRs that carries this sequence. In most classification systems, including binary trees ($N_B = 1$), this relationship is missed because T is found in proteins of two very different groups (the Beta Adrenoceptors and the New Composite Group 2) which are split apart from each other early in the tree due to another motif (Liu and Califano 2003). Another example demonstrates the utility of structure tidying. Here, N_B is set to 1, leading to the splitting of the biological class GLYCHORMONER early on due to a different, highly statistically significant motif that is present in some members of the class but not others. Later in the hierarchy, the two subsets of the GLYCHORMONER family end up as two classes, each of which is associated with a list of motifs. These lists turn out to be very similar, leading these two groups to be merged during structure tidying. In fact, the authors find that using initial values of ($N_B = 1$, N_D = infinity) (basically a binary tree) but then refining using structure tidying allows them to generate classes which stack up very well against a set of "biological" classes obtained from a merging of information from the GPCR database (GPCRDB) and PRINTS (Liu and Califano 2003).

APPLICATION OF CASTOR TO THE OLFACTORY RECEPTOR FAMILY

The utility of motif based functional classification of proteins has been further demonstrated by the recent application of the CASTOR program (described above) to the olfactory receptor (OR) family (Liu *et al* 2003). The OR family is a large family of G protein coupled receptors, which in mammals takes up a surprising 1-3% of the genome (Liu *et al* 2003). 1296 OR genes have been found in the mouse, ~1000 of which are expected to be functional (the remaining genes are probably pseudogenes) (Zhang and Firestein, 2002). A single OR is expressed on each olfactory sensory neuron in the olfactory epithelium (Ressler *et al* 1994 and Vassar *et al* 1994). The ORs bind to odorants, creating a spatial pattern of activity across the olfactory sensory neurons that represents the chemical properties of the odorant. Multiple molecular features may allow an odorant to bind to multiple different ORs, and a given OR may bind to several different odorants that have a chemical property in common (Araneda *et al* 2000, Mori *et al* 1999, Zhao *et al* 1998, Malnic *et al* 1999). The mammalian olfactory system has the ability to detect and discriminate between an astounding array of chemical compounds that may be present in the external

environment (Buck and Axel 1991, Zhang and Firestein 2002, Buck 2000), and it has been suggested that this ability may be due to combinations of activated domains within the ORs (Uchida *et al* 2000).

In their analysis of the OR family, Liu *et al* used 1332 potential full-length intact amino acid sequences from the mouse and human genome databases, covering at least 90% of the known mouse and human olfactory receptor genes (Liu *et al* 2003). CASTOR was used as described above, with the best results (with respect to the few known global motifs for the OR family) obtained using the values ($N_B = 2$, N_D = infinity) followed by class space pruning and structure tidying. With these parameters, CASTOR found 86 motifs with 76 distinct support sets, where the support set of a motif is defined as the group of genes that carries that motif. The support sets ranged in size from 15 to 1330 OR sequences. A motif is defined as "conserved" within a group of proteins if it is present in the vast majority of the members of the group, and is defined as "specific" to the group if it is found in members of the group but virtually nowhere else in the database. A motif is also defined to "characterize" a group if it is present in almost all members of the group and nowhere else in the database.

Ability of motifs found in this analysis to characterize known classes of ORs within the olfactory database: Not surprisingly given the overall similarity of the OR family, 10 of the 86 motifs were found to be conserved in the full OR database; of these, five were found not to be specific to the ORs, as they were also conserved in a database of 846 non-OR GPCRs. In addition, motifs characterizing the previously delineated Class 1 and Class 2 OR subgroups were found, as were motifs that characterize the subset of mouse ORs (Liu *et al* 2003).

Correlations of available functional data with OR groups defined by this analysis (ORs that share combinations of motifs): If each of the 1332 ORs in the database is put into a group based on the specific set of motifs that it carries, the database is subdivided into over 1227 subgroups, 91 of which had more than one member. In addition, the authors constructed larger, non-unique subgroups. Each member of a given one of these subgroups contains all of the ORs that define the subgroup, but may have additional motifs as well. To test the utility of these analyses, subgroups were compared to putative ligand binding groups that had previously been predicted to share function. A set of three motifs was found to characterize 16 total sequences, 8 of which represent 8 out of 12 sequences in a genomic region known to be involved in detection of isovaleric acid. Another set of 15 motifs was found to characterize a subgroup of 38 ORs, 8 of which are suspected based on expression pattern in the olfactory epithelium to be functionally related. Finally, a set of three ORs that had been shown previously to recognize nonanedioic acid all fell within a group of 17 ORs characterized by a combination of four motifs. It is remarkable that these ORs are grouped together by this analysis, because while two of the ORs have 96% identity, the third is only 33% identical to the other two. In phylogenetic classifications, the smallest cluster that included all three of these nonanedioic acid – recognizing ORs is the major Class I cluster, which includes 100 ORs. Additionally, of the 17 ORs in the group, one recognizes nonanoic acid, which is closely related to nonanedioic acid (Liu *et al* 2003).

None of the methods of grouping ORs to date (including phylogenetic analysis, principle component analysis, and grouping based on genomic location) has correlated well with the functional data that is available for ORs (Liu *et al* 2003). The success of CASTOR in grouping

these ORs using a motif based approach suggests that this method may be able to provide biologically relevant groupings of proteins in families where other methods have failed, and may be able to offer additional, complementary insights even in families in which other approaches have succeeded.

PROPOSED APPLICATION OF MOTIF BASED FUNCTIONAL CLASSIFICATION TO VOMERONASAL RECEPTORS

The vomeronasal receptors (Vrs) are a large family of G protein coupled receptors believed to be involved in pheromone detection (Buck 2000, Del Punta *et al* 2002). Vr expressing neurons are found in the epithelium of the vomeronasal organ, and each neuron expresses only one Vr. While in mice there are approximately 1000 olfactory receptors, there are considerably fewer Vrs, probably in the range of 200 (Buck 2000, Lane *et al* 2002). The Vrs can be divided into two main families, Vr1 and Vr2, and no motifs are conserved between the two families or between these families and the olfactory receptors (Del Punta *et al* 2000). Until recently (2000) only 5 sequences of mouse Vr1 receptors were publicly available (Del Punta *et al* 2000), but a recent search of Genbank found that there are now 107 Vr1 protein sequences available (www.ncbi.nlm.nih.gov). These proteins are believed to function similarly to the ORs in that different receptors may recognize different ligands (Buck 2000). Given the success of CASTOR in motif based classification of ORs (Liu *et al* 2003), motif based functional classification might be a very informative approach to apply to the study of sequences within the Vr family. While ideally an automated program such as CASTOR would be used to find motifs and classify all of the V1r and V2r proteins, as a first step CASTOR or a similar program could be applied to the 107 currently available V1r proteins. Since there is very little experimental data available for this family, functional predictions generated based on several combinations of N_B and N_D values could be useful in guiding future experimental work.

REFERENCES

Areneda RC, Kini AD, Firestein S. The molecular receptive range of an odorant receptor. Nature neurosci 3, 1248-55 (2000).

Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**, 175-187 (1991).

Buck, L. B. The molecular architecture of odor and pheromone sensing in mammals. *Cell* 100, 611-618 (2000).

Del Punta K, Rothman A, Rodriguez I, Mombaerts P. Sequence Diversity and Genomic Organization of Vomeronasal Receptor Genes in the Mouse. Genome Research 10, 1958-67 (2000)

Gotoh O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J Mol Biol 264, 823-38 (1996).

Heger A, Holm L. Towards a covering set of protein family motifs. Progress in Biophy and Mol Bio 73 321-37 (2000).

Lane RP, Cutforth T, Axel R, Hood L, Trask BJ. Sequence analysis of mouse vomeronasal receptor gene clusters reveals common promoter motifs and a history of recent expansion. Proc Nat Ac Sci 99, 291-96 (2002).

Liu AH and Califano A. CASTOR: Clustering algorithm for sequence toxonomical organization and relationships. J of Comp Bio 10, 21 – 45 (2003).

Liu AH, Zhang X, Stoovitzky GA, Califano A, Firestein SJ. Morif-based construction of a functional map for mammalian olfactory receptors. Genomics 81, 443-56 (2003).

Malnic, B., Hirono, J., Sato, T. & Buck, L. B. Combinatorial receptor codes for odors. *Cell* **96**, 713-723 (1999).

Mori, K., Nagao, H. & Yoshihara, Y. The olfactory bulb: coding and processing of odor molecule information. *Science* **286**, 711-715 (1999).

Mount DW. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Press, Cold Spring Harbor NY (2001).

Mulder NJ, Apweiler R. Tools and resources for identifying protein famiies, domains and motifs. Genome Biology 3 (2001).

Ressler KJ, Sullivan SL, Buck LB. Information coding int eh olfactory system: evidence for a stereotyped and highly organized epitope map in the olfactory bulb. Cell 79, 1245-55 (1994).

Uchida N, Takahashi YK, Tanifuji M, Mori K. Odor maps in the mammalian olfactory bulb: doman organizatin and odorant structural features. Nature Neruoscience 3, 1035-43 (2000).

Vassar R, Chao SK, Sitcheran R, Nunez JM, Vosshall LB, Axel R. Topographic organization of sensory projections to the olfactory bulb. Cell 79, 981-91 (1994).

Zhao H, Ivic L, Otaki JM, Hashimoto M, Mikoshiba K, Firestein S. Functional expression of a mammalian odorant receptor. Science 279, 237-42 (1998).

Zhang, X., & Firestein, S. The olfactory receptor gene superfamily of the mouse. *Nature Neurosci* 5, 124-33 (2002).