PROTEINS: Structure, Function, and Genetics 53:683–692 (2003)

Automatic Annotation of Protein Function Based on Family Identification

Federico Abascal* **and Alfonso Valencia**
Protein Design Group, National Centre for Biotechnology, CNB-CSIC, Cantoblanco, Madrid, Spain

ABSTRACT **Although genomes are being sequenced at an impressive rate, the information generated tells us little about protein function, which is slow to characterize by traditional methods. Automatic protein function annotation based on computational methods has alleviated this imbalance. The most powerful current approach for inferring the function of new proteins is by studying the annotations of their homologues, since their common origin is assumed to be reflected in their structure and function. Unfortunately, as proteins evolve they acquire new functions, so annotation based on homology must be carried out in the context of orthologues or subfamilies. Evolution adds new complications through domain shuffling: homology (or orthology) frequently corresponds to domains rather than complete proteins. Moreover, the function of a protein may be seen as the result of combining the functions of its domains. Additionally, automatic annotation has to deal with problems related to the annotations in the databases: errors (which are likely to be propagated), inconsistencies, or different degrees of function specification. We describe a method that addresses these difficulties for the annotation of protein function. Sequence relationships are detected and measured to obtain a *map* of the sequence space, which is searched for differentiated groups of proteins (similar to islands on the *map*), which are expected to have a common function and correspond to groups of orthologues or subfamilies. This mapmaking is done by applying a clustering algorithm based on *Normalized cuts* in graphs. The domain problem is addressed in a simple way: pairwise local alignments are analyzed to determine the extent to which they cover the entire sequence lengths of the two proteins. This analysis determines both what homologues are preferred for functional inheritance and the level of confidence of the annotation. To alleviate the problems associated with database annotations, the information on all the homologues that are grouped together with the query protein are taken into account to select the most representative functional descriptors. This method has been applied for the annotation of the genome of *Buchnera aphidicola* (specific host *Baizongia pistaciae*). Human inspection of the annotations allowed an estimation of accuracy of 94%; the different kinds of error that may appear when using this approach are described. Results can be accessed at http://www.pdg.cnb.uam.es/funcut.html. The programs are available upon request, although installation in other systems may be complicated. Proteins 2003;53:683–692.** © 2003 Wiley-Liss, Inc.

Key words: protein function prediction; genome analysis; protein families and subfamilies; orthologues and paralogues; annotation inconsistencies; database errors

INTRODUCTION

Spectacular progress has been made in the automation of experimental techniques in molecular biology, especially those for genome sequencing, functional genomics, and proteomics. In all cases, however, it is difficult to reach valid biological conclusions based on the massive data generated by such approaches and the development of computational methods is still a bottleneck. The two key steps in the analysis of genomic data are the identification of the genes in the raw DNA sequence and the prediction of the function of the corresponding open reading frames (ORFs). This study focuses on the second step.

THE PROCESS OF FUNCTION PREDICTION BY TRANSFERENCE OF INFORMATION FROM HOMOLOGOUS SEQUENCES

The first step is the search for similar sequences in databases, followed by the selection of homologous sequences from the set of similarities, i.e., identifying sequences with a common evolutionary origin. The statistics of sequence similarities[1] are commonly used for this purpose. Unfortunately, identification of the homologues is not sufficient to guarantee a correct transference of functional annotation.[2–6] In the course of evolution, homologues differentiate to embrace new functions, corresponding to the organization of sequence families and subfamilies.[7] For example, the superfamily "P-loop containing nucleotide triphosphate hydrolases" includes families as varied as "RNA helicases," "G proteins," and "ABC transporters." Moreover, in the "G proteins," families special-

Grant sponsor: Spanish Ministry of Science and Technology (CICYT);
Grant sponsor: Government of Madrid.

*Correspondence to: Federico Abascal, Protein Design Group, National Centre for Biotechnology, CNB-CSIC, Cantoblanco, Madrid E-28049, Spain. E-mail: fabascal@cnb.uam.es

Received 14 January 2003; Accepted 25 February 2003

ized in various cell functions co-exist, such as ras-related proteins, involved in the cell cycle; *rab*, related to vesicle cell traffic; *arf*, also part of the protein trafficking machinery; elongation factors Tu and G, and others. The process of transference of functional annotation based on the identification of homologues requires a deeper analysis of the structure of the corresponding families and their associated functions.

A complete reconstruction of the phylogenetic tree of the family could be required for the analysis of the relation between family groups and functions. Unfortunately, the process of tree building and family analysis is difficult to automate.

Function-Prediction Errors Introduced by "Classical" Annotation Strategies

Most of the current systems for automatic annotation obviate the step of analyzing homologies in terms of families and subfamilies and directly transfer function from the most similar sequence identified in the databases, for example, by selecting the best hit after searching with standard tools such as BLAST.[8] The obvious difficulty of encapsulating in a single relation all the functional information associated with a given protein family makes this annotation process a poor substitute for the deeper analysis of the structure of the protein family structure. There is copious literature on the errors introduced by the iteration of this process (see for example Brenner[9] and Devos and Valencia[10]). Indeed, the systematic comparison of the annotations stored in databases has made it clear that a significant number of homologous sequences have different functions.[7,11]

The correct characterization of the relation between the sequences forming a protein family may solve some of these problems, even if the final process of annotation still relies on the existing database annotations. Hence, such characterization could, therefore, be subject to the errors introduced during this process. Efforts are underway to improve the database annotations by keeping pointers to the origin of the annotations. An attempt is also being made to create systems that facilitate this process by automatically retrieving information from the original textual sources.[12]

Domain Shuffling and Other Difficulties

The process of classifying protein families into the corresponding families and subfamilies is not trivial. The study of the evolutionary relationships between proteins is often complicated by the presence of multiple domains. Domains can have different origins, and be associated with several domains in different proteins. Moreover, the function of a protein can be seen as the result of combining the functions of its domains. Therefore, for any function prediction schema it is essential to take into account the domain structure of proteins and avoid transferences based on incomplete domain identification.

After identifying the sequence relations in the corresponding families, however, further difficulties arise associated with the transference of the corresponding annotations. It is quite common to find annotations of different natures or ones addressing a different level of detail for a given function. For example, of two proteins belonging to the ras-p21 subfamily, Q9BJ57 is annotated as "Ras" in TrEMBL, while RASH_HUMAN is annotated as "Transforming protein P21/H-RAS-1 (C-H-RAS)" in Swiss-Prot, with a deeper definition of the protein function. An example of different degrees of specificity may be found in the case of PS2_HUMAN and PS2_MOUSE, annotated as "PS2 protein precursor (HP1.A) (Breast cancer estrogen-inducible protein) (PNR-2)" and "PS2 protein precursor," respectively. In the first case, given the importance of the data that relate this protein to human cancer, this specific information is included in the description of the function. Therefore, the analysis of the annotations requires a balance between general descriptions and detailed ones, corresponding to higher accuracy but less information in the case of the general ones.

RELATED APPROACHES TO THE AUTOMATIC PREDICTION OF PROTEIN FUNCTION BASED ON THE INFORMATION OF HOMOLOGOUS SEQUENCES

Some of the recent approaches are introduced in the following section.

GeneQuiz[13] was the first system that completely automated the task of sequence analysis and function annotation. The functional information was obtained from the potentially most informative sequence selected among the set of similar ones. Rules for the selection of informative sequences were related to the database of origin, presence of fragments, and type of keywords associated to them. Genequiz delivered annotations with an associated confidence value that was related with the similarity (fixed e-value) with the sequence selected for the information transfer. The system included a simplified lexical analysis for identifying the informative descriptions (it has inspired the lexical analysis that we apply).

EditToTrembl[14] is a system in which the annotation of sequences is based on the intricate execution of different programs to predict such features of the proteins as transmembrane regions, subcellular location, or enzymatic codes. A complement of this work is that of Fleischmann et al.,[15] where a method for automatic functional annotation is described. Thus, these authors overcome some of the previously mentioned difficulties, such as the functional transfer from the best hit. The method is based on the use of PROSITE[16] as an external database to cluster the sequences of a reference database (Swiss-Prot[17] in this case) into groups. Each group is inspected to identify functional information shared by most of the proteins. In the positive cases, an annotation rule is derived for the annotation of new sequences. Additionally, some rules are applied to reduce the potential number of false positives, for example, the taxonomy of the query sequence must match the species distribution of the proteins described by the (PROSITE) condition. This conservative approach is currently applied during the generation of the TrEMBL entries, producing an enrichment of the

functional annotations prior to the more precise annotation work carried out by the Swiss-Prot curators. Using the PROSITE groups as seeds of the process imposes limitations for the coverage of the annotations.

The PRECIS system[18] is more a distiller of information than an annotation method. Basically, it receives a set of Swiss-Prot identifiers (which could come, for example, from a BLAST report) and distils a functional report combining their annotations, thus removing redundancies and applying some rules and filters for the different fields of the Swiss-Prot entries.

Andrade[6] described a method for addressing the problem of the annotations specific to protein domains by using position-specific annotation of protein function based on the analysis of multiple homologous sequences. The multiple sequence alignment corresponding to the homologues was used to assign functions to specific domains (positions covered by the sequences in the alignment). The functional descriptions of the similar sequences were processed and screened for common strings of words. The correlation between the conservation of the aligned positions with respect to the query protein and the presence of common functional descriptors in the aligned proteins was used to produce annotations common to the shared positions. This procedure allowed the construction of consensus descriptions attached to defined regions of the query protein. The two main difficulties of this idea are perhaps the complexity of its automation and the fact that it will work properly only under certain conditions (for example, it requires that the set of homologues contain proteins belonging to the same subfamily of the query).

The approach presented here tries to overcome many of the aforementioned obstacles for function annotation. It applies a clustering algorithm to classify proteins into families and subfamilies. A lexical analysis inspired on GeneQuiz is applied to identify informative functional descriptions. As in Andrade[6] and Fleischmann et al.,[15] our approach uses information from multiple homologues to select the information to be transferred. The problem of transferring functions from unrelated domains is addressed by analyzing the degree of coverage of the corresponding local sequences alignments.

METHODS

Algorithm for the Assignment of Functional Annotations Based on the Analysis of Sequence Clusters

The workflow of the method proceeds as follows:

1. A sequence similarity search is carried out to find proteins related to the query sequence. A clustering algorithm is applied in order to identify closely related sequence groups in the set of similar proteins. More related sequences are more likely to share a common function. In some cases, recursive sequence similarity searches lead to better representation of the related subfamilies, which facilitates the clustering.
2. The local alignments with the closely related proteins clustered together with the query protein are classified in different *categories* depending on the extent to which the alignments cover the length of the query and target sequences (*alignment categories*).
3. Key functional annotations of the corresponding proteins are analyzed, including functional descriptions, enzymatic activity codes, and Swiss-Prot style keywords.
4. The transference of information is carried out starting from the alignment categories with a better coverage. A confidence level is assigned to each one of the annotations. This level is derived from the alignment categories.

Step 1. Sequence searches

The similarity searches were carried out with BLAST on a non-redundant database (nrdb program from NCBI at ftp://ftp.ncbi.nih.gov/pub/nrdb and cd-hit[19]) that included Swiss-Prot, TrEMBL, and TrEMBLnew. Those sequences with a similarity value above a cut-off (e.g., E-value < 0.1) were further BLAST-aligned between them to obtain a rough measure of their pairwise similarity (i.e., their E-values). This procedure effectively maps the sequence space surrounding a query sequence. If desired, this local sequence space map can be extended to other more divergent related subfamilies through iterative intermediate sequence searches,[20,21] exploiting the transitivity principle of homology: if protein A is homologous to B, and B to C, then A and C are homologues (if the domains shared in the A-B and B-C relationships correspond to each other). Moreover, recursive searches provide better descriptions of the sequence space, so the clustering works better. The use of PSI-BLAST or other profile-based methods is inadequate for this classification process because PSI-BLAST does not return measures of the *distance* between pairs of proteins but *distance* between proteins and profiles. A post-processing of the PSI-BLAST results by, for example, realigning all the results, could be a valid alternative.

Identification of subfamilies by clustering. The complete set of distances between the sequences provides the basic description of the local sequence space. A clustering process is used to identify groups of sequences that more likely correspond to protein subfamilies. The algorithm used is the "Normalized Cut,"[22] and the application of this graph theory to biological sequences is described in Abascal and Valencia.[23] A weighted undirected graph G(V, E) represents the sequence space. The *nodes* (V, sequences in this case) are connected through *arcs* (E) that represent their similarity relationship. Each arc has an associated *weight* (w), proportional to the similarity measured between the sequences in the form of $-\log_{10}$(E-value) of the BLAST algorithm. A cut (A, B) in a graph is a partition of G into two sets of nodes A and B, obtained by removing some of the arcs. The capacity of a cut is the sum of the weights of the arcs that have to be removed to obtain the cut (A,B). The *minimum cut*[24] of a graph is the one with minimum capacity. The minimum cut provides an effective measure of the separation of the initial sequence space in two sequence groups.

$$\text{cut(A,B)} = \text{Sum w(i,j); i in A, j in B}$$

Shi and Malik[22] proposed a *normalization* of the capacity of the cuts by including in the formula the amount of connections of each one of the two separate sequence groups:

$$Ncut(A,B) = cut(A,B)/asso(A,V) + cut(A,B)/asso(B,V)$$

where asso(A,V) is the sum of the weights of the arcs from all nodes in A to all nodes in V (including those in A).

Once the best cut is determined, the algorithm proceeds recursively, searching for new cuts of the established sequence groups. The recursive process continues until neither of these two conditions holds:

1. The arithmetical mean of the arc weights in A or B exceeds by twofold the arithmetical mean of the arc weights that cross from A to B.
2. The number of arcs divided by the maximal possible number of arcs is higher in A or B than the same measure in G.

An evaluation of this clustering method applied to protein sequences was carried out in Abascal and Valencia.[23] The outcomes of the clustering process are a number of groups of well-defined and highly connected sequences and the *distances* between these clusters. One of these groups is the one containing the query protein and ideally will contain the other members of its subfamily.

Step 2. Analysis of the Local Alignments and Assignment of "Alignment Categories"

The local alignments of the query sequence versus the closely related proteins are analyzed to determine the degree to which they cover the corresponding sequence lengths. We have defined four categories:

1. Both the query and the target align through more than 80% of their lengths. In this case, the functional transfer is considered to be secure and complete.
2. The entire length of the query sequence cannot be aligned with the target sequence. In this case, the transference of functional annotations might not be complete since the query protein may contain differential properties associated to the region of the sequence not matched by the target protein.
3. Less than the entire length of the target protein is aligned with the query protein. In this case, the transference of annotations could be wrong if some of the functions of the target protein are associated to the sequence region not aligned with the query sequence.
4. The less confident category corresponds to the cases in which neither the query nor the target align completely.

In the cases where the target protein is annotated as FRAGMENT in the database, its alignment is always taken as incomplete.

Step 3. Definition of the Functional Annotations to Be Transferred

The goal of functional descriptions is to find descriptions of function that can be compared with the information provided in the standard "DE" field of Swiss-Prot. Having a set of proteins belonging to the same subfamily (or group of orthologues) facilitates the secure transfer of a good functional description, since we can select the most representative description (the most homogeneous compared to the co-clustered sequences), as follows:

1. *Noninformative word filtering*. Descriptions are filtered to remove words that contain no information about function, such as FRAGMENT, HYPOTHETICAL, or COSMID.
2. *Deriving a "homogeneity" score for each description to measure how representative it is.* Each description is split into its component words, and for each word the frequency with which these words appear in the set of descriptions in the cluster is calculated. A pre-score is calculated for each description by adding together the frequencies of its words. This pre-score is divided by a correction factor to avoid the bias towards long descriptions. This factor is defined as the number of words divided by the number of synonyms (that usually are very informative and are given between parenthesis in the Swiss-Prot entries). A *normalized* homogeneity score is calculated as the fraction that each description score represents in the sum of all the description scores.
3. *Weighting normalized homogeneity scores with normalized similarity scores.* Similarity scores are *normalized* in the same manner, by calculating the fraction each BLAST similarity score represents in the total sum of BLAST similarity scores. Both normalized scores are weighted in a *combined score*, what is useful for the cases where two or more subfamilies erroneously get clustered together.

As will be explained later, alignment categories and these combined description scores are used to select the annotation to be transferred. Once selected, the description is inspected to reject descriptions that contain no information. The process of identification of non-informative descriptions is based on lexical analysis, inspired in that of GeneQuiz.

Examples of frequent non-informative annotations are "[Hypothetical|Putative] [Mol.Weight] [Lipo|Glyco]Protein [word]" where, if "word" is present, it should be present in at least some of the other descriptions of the proteins in the cluster to avoid rejection."[Mol.Weight]" represents in perl \d+(\.)*(\d)*(\s)*K(D)*(A)*(\s)*. The character "|" means "or." Words inside "[ ]" may or may not appear, etc.

Some of the words commonly found in non-informative descriptions are: "Intergenic", "Cosmid" or "Genomic sequence," their presence is enough to label descriptions as non-informative. Another rule to identify uninformative descriptions is to remove all (expected) uninformative words and check for the remaining words if they are present in at least some of the others descriptions in the cluster.

Finally, informative descriptions are cleaned by removing words that frequently appear in functional descriptions but are not transferable based on homology, such as the molecular weight or the word "fragment."

Proteins: keywords	Keywords: counts
Prot 1: A B C	KW counts:
Prot 2: A B C	A: 8
Prot 3: A B C	B: 5
Prot 4: A B D	C: 3
Prot 5: A B D	E: 3
Prot 6: A E F	F: 3
Prot 7: A E F	D: 2
Prot 8: A E F	

Fig. 1. Key words accepted: A B C. Note that C, E, and F have the same frequency, but only C is transferred to avoid mixing key words that do not co-occur. The process: first A is selected as seed. Then, since B is connected to A more than 4 times (8/2), B will also be accepted. Then C will also be accepted because it is connected to B more than 2.5 times (5/2). No more keywords will be added.

Enzymatic codes. Since EC numbers are expressed in a non-ambiguous language, there is no need to measure their *homogeneity*. The EC number is transferred to the query protein from the sequence inside its sequence cluster with the highest similarity score and a better alignment category.

Key words. Functional key words assigned in Swiss-Prot depend strongly on the functional domain organization of the proteins, e.g., Myristate, Calcium-binding. We have preferred to transfer only key words in the cases where the target protein is completely covered by the alignment (first and second alignment categories). The key words frequency is calculated, and a graph is built using key words as nodes, and the arcs connecting the nodes are labelled by the number of times that key words appear associated to the same protein. The selection process is based on the co-occurrence of the key words, and it is applied reiteratively to rescue partial co-occurrences. First, a key word score is calculated for each protein by weighting similarity and homogeneity scores, as in the case of the functional descriptions). The key word with the highest frequency selected among the ones of the protein with the highest key word score is accepted, and selected as seed. Repeated searches recover other key words connected to some of the accepted ones with at least the half of its frequency. The frequency must be half or greater to avoid including key words that are not co-occurring, as illustrated in Figure 1.

Neighbor clusters annotation. We have incorporated an additional procedure to extract information from neighboring sequence clusters. In this case, the intention is to provide general annotations for each of the sequence clusters. This procedure is particularly helpful in those cases where the cluster of sequences around the query protein does not contain enough proteins with relevant functional annotations.

Essentially, the procedure works by selecting the word that is most frequent in the set of descriptions and all the other words that are frequently associated to it. The position of these words with respect to the most frequent one is used to order them and to build the final functional cluster description. In detail, the steps are:

- For each of the neighbor clusters its own set of annotations is retrieved.
- For each protein description, uninformative words are removed and the frequency of the remaining ones is measured.
- We build a graph in which there is one node for each of the words and the (weighted) arcs between the nodes reflect the number of descriptions in which two words co-occur.
- The most frequent word is selected as *seed,* and the graph is searched for all those nodes (words) connected to its node with a frequency at least half of the *seed* frequency (slightly different from the approach used with the keywords).
- Then, for each of those *accepted* words, the most frequent position where they appear relative to the seed word is determined.
- Finally, the description is built by sorting the list of words according to these most frequent positions. Even if this sorting procedure is not perfect, it is simple enough to give an idea of the functions that are present in the neighbor protein clusters.

Step 4. Assignment of the Functional Annotations to the Different Alignment Categories

For the final transfer of functional description, the proteins are inspected from the best alignment coverage category to the worst. In each of the categories, the best description (higher combined score) is searched for. If there are no descriptions in that category, or if the best one is considered noninformative, we go down to the following category and search again for the best description in this category. The confidence of the transference is derived from the alignment category.

RESULTS

Selected Examples of Functional Annotation

The parameters used for the recursive sequence similarity searches have been selected to obtain clearer results in each of the examples. They are different in each case because of the different sizes of the different families analyzed. For the case of Buchnera's genome annotation, we did not use recursive searches but single BLAST searches with realignment of all the results (see Methods).

TABLE I. BLAST's Best Hits for Swiss: TETM_NEIME†

Sequences producing significant alignments:	Score (bits)	E value
TET1_ENTFA (Q47810) Tetracycline resistance protein tetM from tr. . .	1225	0.0
TETS_LACLA (Q48712) Tetracycline resistance protein tetS (Tet(S)).	981	0.0
TETO_CAMCO (P23835) Tetracycline resistance protein tetO (Tet (O)).	979	0.0
TETW_BUTFI (O52836) Tetracycline resistance protein tetW (Tet(W)).	885	0.0
Q93K56 (Q93K56) Tetracycline resistance protein.	858	0.0
Q9RLW0 (Q9RLW0) TetT.	558	e-158
P70882 (P70882) TETA (Q) 3 PROTEIN.	501	e-141
TETP_CLOPE (Q46306) Tetracycline resistance protein tetP (Tetb(P)).	437	e-122
Q97J38 (Q97J38) Tetracycline resistance protein, tetQ family, GT.	426	e-118
TETM_STRLI (Q02652) Tetracycline resistance protein tetM.	326	3e-88
AAK87139 (AAK87139) AGR_C_2489p.	317	2e-85
OTRA_STRRM (Q55002) Oxytetracycline resistance protein.	308	7e-83
Q97KR3 (Q97KR3) Tetracycline resistance protein tetP, contain GT.	263	2e-69
Q8XLR6 (Q8XLR6) Probable tetracycline resistant protein.	257	1e-67
EFG_THETH (P13551) Elongation factor G (EF-G).	256	2e-67
Q9AIG7 (Q9AIG7) Elongation factor G.	251	1e-65
EFG_AQUAE (O66428) Elongation factor G (EF-G).	251	1e-65
EFG_THEMA (P38525) Elongation factor G (EF-G).	247	2e-64
Q8YP62 (Q8YP62) Translation elongation factor EF-G.	243	3e-63
Q9PI16 (Q9PI16) Elongation factor G.	242	6e-63
EFG_CHLMU (Q9PJV6) Elongation factor G (EF-G).	241	1e-62
BAB56709 (BAB56709) Translational elongation factor G.	240	2e-62
Q9F4B2 (Q9F4B2) Translational elongation factor G, EF-G (Fragment).	239	3e-62
EFG_SYNP6 (P18667) Elongation factor G (EF-G).	239	3e-62
Q9RXK5 (Q9RXK5) ELONGATION FACTOR G.	238	8e-62

†It can be appreciated that BLAST e-values order appropriately the sequences of the *tet* and EF-G subfamilies. Even if there's not a clear separation attending to the magnitude of the e-values, the clustering algorithm distinguishes both subfamilies, but fails to include two more divergent *tet's* in the proper cluster. Proteins TET1_ENTFA to OTRA_STRRM belong to the same group; proteins Q97KR3 and Q8XLR6 form another cluster; the remaining ones form the third group. The complete BLAST result can be obtained from: http://www.pdg.cnb.uam.es/fabascal/SEARCH AND CLUS/TETM NEIME/Q51238.bls.

TABLE II. Subfamilies Found by the Recursive Searches for Swiss::TETM_NEIME and the Subsequent Clustering, Which Resulted in 21 Clusters

Cluster id.	Size	Subfamily
2	59	Translation initiation factor IF-2
3	50	GTP-binding protein IepA
4	35	GTP-binding protein TypA/BipA
5	74	Elongation factor 2 (EF-2)
10	80	Elongation factor G (EF-G)
13	13	Tetracycline resistance protein tet[W M S R . . .]
14	24	Peptide chain release factor 3 (RF-3) *(bacteria)*
15	248	Elongation factor 1-alpha plus 18 *Eukaryotic* peptide chain release factor 3
17	25	NodQ bifunctional enzyme and CysN/cysC bifunctional enzyme
18	11	Selenocysteine-specific elongation factor
21	117	Elongation factor Tu (EF-Tu)

Those containing more than 2 sequences are represented. Note that some subfamilies may be incomplete because similarity searches were limited to a maximum of 750.

TETM_NEIME

The first selected example corresponds to the tetracycline resistance protein, (*tetM)* from *Neisseria meningitidis*. The BLAST search with this protein vs. a nonredundant database selected at a 90% identity level rendered the results shown in Table I. In this case, there is not a clear separation of the subfamilies of TET and EF-G proteins based on E-values. However, the clustering of the sequence space local to the query protein (obtained by applying three rounds of *intermediate sequence searches* with a cut-off E-value of 1e-07), allowed the appropriate separation of the two subfamilies (Table II) and also the correct classification of other more distant subfamilies. Assuming that the co-clustered sequences share a common function makes it possible to use them as sources of annotation, analyzing the descriptions as described in Methods. This yields the annotation for the query protein: TETRACYCLINE RESISTANCE PROTEIN TETS

(TET(S)) instead of TETM. This is an especially problematic case where the clustering is not able to classify into separate groups the different tetracycline resistance determinants. Instead, it puts them together according to their high similarity. The annotations in the database (or the nomenclature) seem to be inconsistent (or the specificity has no evolutionary foundation), because the percentage of sequence identity is much higher between some Tet(M) and Tet(whatever) than between two Tet(M), for example, TETM_NEIME vs. TETS_LACLA (77%) and TETM_NEIME vs. TETM_STRLI (35%). A phylogenetic tree of the best BLAST hits of TETM_NEIME can be see at http://www.pdg.cnb.uam.es/funcut.html. The keywords for the co-clustered sequences were:

Q02652 Protein biosynthesis; Antibiotic resistance; GTP-binding.
Q93K56 GTP-binding.
Q46306 Protein biosynthesis; Antibiotic resistance; GTP-binding.
Q51238 Protein biosynthesis; Antibiotic resistance; GTP-binding; Plasmid.
Q47810 Protein biosynthesis; Antibiotic resistance; GTP-binding; Transposable element.
P23835 Protein biosynthesis; Antibiotic resistance; GTP-binding.
Q97J38 Complete proteome.
O52836 Protein biosynthesis; Antibiotic resistance; GTP-binding.
Q48712 Protein biosynthesis; Antibiotic resistance; GTP-binding; Plasmid.
P70882 GTP-binding.
Q55002 Protein biosynthesis; Antibiotic resistance; GTP-binding.
Q9RLW0 GTP-binding.

from which were derived the following key word annotations: GTP-binding; Protein biosynthesis; Antibiotic resistance.

The neighbor clusters and their annotations were:

ID:10; SIZE:80; PROXIMITY:46.95 *ELONGATION FACTOR G EF*
ID:14; SIZE:24; PROXIMITY:26.95 *PEPTIDE CHAIN RELEASE FACTOR 3*
ID:4; SIZE:35; PROXIMITY:20.42 *GTP BINDING PROTEIN TYPA*
ID:3; SIZE:50; PROXIMITY:16.25 *GTP BINDING PROTEIN LEPA*
ID:5; SIZE:74; PROXIMITY:13.45 *ELONGATION FACTOR 2*
ID:21; SIZE:117; PROXIMITY:10.51 *ELONGATION FACTOR TU EF*
ID:2; SIZE:59; PROXIMITY:6.13 *TRANSLATION INITIATION FACTOR IF 2*
ID:15; SIZE:248; PROXIMITY:3.07 *ELONGATION FACTOR 1 ALPHA*
ID:17; SIZE:25; PROXIMITY:1.96 *SULFATE ADENYLYLTRANSFERASE SUBUNIT 1*

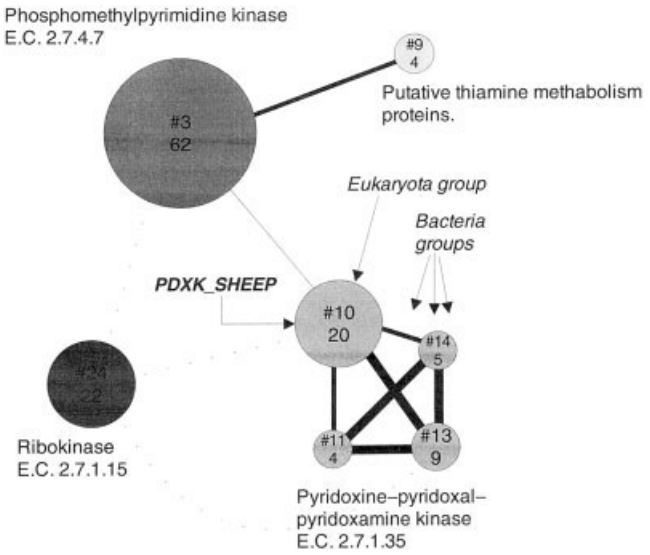

Fig. 2. Note that some subfamilies may be incomplete because recursive searches were stopped before convergence. Each circle and its radius correspond with a cluster and its size. The numbers inside indicate the cluster **id** and the number of proteins in it. The width of the lines connecting the clusters represents the strength of their connection. The different gray intensities correspond to the different families.

EC 2.7.7.4 ADENYLATE TRANSFERASE SAT ATP SULFURYLASE LARGE
ID:18; SIZE:11; PROXIMITY:1.87 *SELENOCYSTEINE SPECIFIC ELONGATION FACTOR SELB TRANSLATION*

These automatic annotations are difficult to read in some cases because of the absence of punctuation characters, such as, for example, the parentheses, which are not managed. Moreover, when a word appears more than once in a description, only the first is counted. For example, the cluster annotation SELENOCYSTEINE SPECIFIC ELONGATION FACTOR SELB TRANSLATION should be SELENOCYSTEINE-SPECIFIC ELONGATION FACTOR (SELB TRANSLATION FACTOR).

One selenocysteine-specific elongation factor is separated from no. 18 to the singleton cluster no. 20. If this "solitary" protein were the query protein, then annotation would not be possible. This represents an example of what kind of errors may appear. Results can be accessed at: http://www.pdg.cnb.uam.es/funcut.html.

PDXK_SHEEP
http://www.pdg.cnb.uam.es/funcut.html

This protein is a PYRIDOXINE KINASE. Recursively searching with it (three rounds with and Evalue cut-off of 1e-03, in this case vs. a 100% non-redundant database comprising Swiss-Prot, TrEMBL and TrEMBLnew), and subsequently clustering the results, yields 29 clusters, corresponding to 160 sequences (first round: 1 sequence; second: 70; third: 89). Of these 29 groups, 7 contain more than three sequences (Fig. 2); the rest correspond to out-layers or cases where the clustering fails to keep the proteins in their corresponding families (a phylogenetic tree is available at the web site).

Annotation. The subfamily of the PDXK_SHEEP is separated in an only eukarya cluster, together with 19 relatives. The bacteria Pyridoxal/pyridoxine/pyridoxamine kinases (including PDXY and PDXK) are divided in three next neighbor clusters. The query protein was annotated with the highest level of confidence as PYRIDOXINE KINASE (PYRIDOXAL KINASE), with enzymatic activity 2.7.1.35 and key words Kinase and Transferase. The other key word in the Swiss-Prot entry corresponding to PDXK_SHEEP, "Acetylation," was not transferred because PDXK_SHEEP was the only protein in the cluster with that key word assigned to it.

The neighbor clusters are annotated as:

ID:13; SIZE:9; PROXIMITY:19.35 *PYRIDOXAMINE KINASE EC 2.7.1.35 PM*
ID:11; SIZE:4; PROXIMITY:16.63 *KINASE*
ID:14; SIZE:5; PROXIMITY:13.78 *PYRIDOXINE KINASE EC 2.7.1.35 PYRIDOXAL VITAMIN B6 PYRIDOXAMINE PN PL PM*
ID:3; SIZE:62; PROXIMITY:1.13 *PHOSPHOMETHYLPYRIMIDINE KINASE*
ID:24; SIZE:22; PROXIMITY:0.69 *RIBOKINASE*

With other more permissive searching parameters (e.g., higher e-value cut-off and additional rounds of sequence searches), other remotely related subfamilies were identified, such as: tagatose-6-phosphate kinase, phosphofructokinases, 2-dehydro-3-deoxygluconokinase, guanosine kinase, adenosine kinase, etc. The analysis of those results involving more clusters and isolated sequences (singletons) was not followed here.

Application to the Annotation of the Buchnera Genome

This annotation method has been applied for the analysis of the genome of *Buchnera aphidicola (*specific host *Baizongia pistaciae*).[25] In this case, the sequence space local to each of the 507 coding genes of buchnera was built with single BLAST searches (e-value cut-off: 0.1) and all vs. all pairwise alignment of the results. The resulting automatic annotations are available at the *funcut* web site. A comparison of automatic annotation vs. BLAST best hit annotations is also available at the same site.

Annotations were manually analyzed in the cases in which no automatic annotation was produced and in the cases of conflict with the annotation of other closely related genomes in the database. This allowed us to obtain an approximate measure of the accuracy of this approach (see Table III).The accuracy was estimated at 94% and three kinds of errors were established for the remaining 6%. Basically, these errors come from unsatisfactory clustering that in some cases divides a given subfamily. This then creates singletons, and in other cases fails to separate two (or more) subfamilies. Other "errors" are due to special characteristics of the lifestyle of this obligate endosymbiont.

1. Singleton errors (21 cases): *fliH, fliJ, fliK, fliM, flgB, flgM* (flagellar proteins). These proteins are clustered separately from their homologues, in singletons, so no functional annotation transfer can be carried out. In Tamas et al. 2002,[26] this divergence for flagellar proteins is also observed and it is proposed that it may be related to the acquisition of new functionalities, since flagelles have not been observed in this bacteria.
2. Too specific descriptions (9 cases): in this case, some specie-specific (not transferable) word in the description is transferred. For example, for *ycfC,* the automatic annotation "Hypothetical protein ycfC (ORF-23)" was manually corrected to "Hypothetical protein ycfC". The word "ORF-23" is particular for the source specie.
3. Incorrect functional assignments (2 cases): Protein *hscA* corresponds to "chaperone protein hscA homologue" but it was automatically annotated as "chaperone protein dnaK" because the clustering did not separate these two very close subfamilies and in the cluster dnaK proteins are much more abundant.

TABLE III. *Buchnera aphidicola* Automatic Protein Function Annotation

Proteome annotation statistics	
Proteome size	507
Correct function assignments	475 (94%)
Errors	32 (6%)
Singleton errors	21
Too specific descriptions	9
Incorrect function	2

An illustrative example of the usefulness of analyzing whether or not the alignments cover the entire sequence length of the involved proteins is the one of *polA*. The best BLAST hits for this protein are DNA POLYMERASES I. In the BLAST similarity list there are also proteins annotated as "Probable 5′-3′ exonucleases." This buchnera protein has lost most of its domains, so alignments with DNA POLYMERASE targets cover 90–30% of query and target lengths. However, alignment with less similar 5′→3′ exonuclease covers 91–97%. This annotation, which seems to be the correct one, was detected automatically.

Keywords and EC Numbers

For the 507 proteins in buchnera's genome, there were 281 EC number assignments, corresponding to 269 proteins (some proteins have more than one enzymatic activity). In the case of key words, 1,463 were assigned to 470 proteins, but if we discount the frequent (but not adequate for transfer) key word "Complete proteome," we have 1,071 key words for 391 proteins. Rejection of nontransferable key words is to be implemented.

DISCUSSION

We have presented a method for the generation of functional annotations based on the study of the annotations of homologous sequences. The method includes new features related to the specific identification of protein subfamilies (orthologous groups) because at this level the function of the homologous proteins tends to be more

conserved than in general protein families (mixture of paralogous and orthologous sequences).

Our method seems to produce correct annotations including those of the "DE" and "KW" fields of Swiss-Prot and enzyme classification numbers (Enzyme Commission code, EC). It is obvious that these three features do not account for all the possibilities of protein function description, and other database annotations are also important for a complete description of function.

It is important to keep in mind that a description of protein function can be done at very different levels from biochemical to cellular. It is appealing to think that the levels more directly related to the chemical function will tend to be more conserved at large sequence distance. Cellular functions, then, which are more dependent on the cellular context and interactions, will be less conserved at the sequence level. Part of this complexity in function description is quite obvious in the comment ("CC") and feature ("FT") Swiss-Prot fields, which include information as varied as catalytic activity, quaternary structure, signal sequences, catalytic residues, domain structure, and post-translational modifications.

The current efforts for the construction of classifications (ontologies) for the definition of function at various levels (particularly the GO Consortium[27]) represent a possible way of alleviating these problems. The use of the current GO ontology in our (and other) automatic annotation methods will not be easy until a substantial number of sequences from various genomes are annotated, something that had still not happened at the time these systems were built, despite the considerable effort made by the database teams.

Selecting Representative Descriptions

Our purpose was to identify the most representative description from a set of functionally related ones. We tested various ways of measuring homogeneity of the descriptions. Both Shannon entropy and the sum of log-probabilities (probabilities understood as the frequency of words) tend to give better scores to long descriptions, even if they contain poorly represented words. The comparison of the frequency of words in a given description with the frequency in the whole set of descriptions using Relative Entropy did not render useful descriptions. In our hands, the best results were obtained by calculating the information content of the description weighted by the number of words in the description. We additionally apply a correction to avoid penalizing longer descriptions that include synonyms.

Prediction of Function from Homologous Sequences and Alternative Approaches

Interestingly, the annotation of function by transference from proteins of related sequences is not the only possibility for the "in silico" prediction of function. The flourishing of genomic data has enabled other modes of function prediction independent of the identification of homologous sequences. The function of proteins can be inferred from the study of the similarity of their expression pattern with those of other genes of known function. Function can also be predicted by exploring the set of interactions deduced by different experimental or computational techniques (for a review see, Valencia and Pazos[28]). According to this idea, different attempts have been made to use the genomic context to improve annotations, for example, by increasing the probability of associating a function with a given protein if it were the best candidate in a given genome.[3,29] Even if interesting attempts to unify the various sources of information on association between proteins and genes are underway (for an early study see Marcotte et al.[30]), problems of consistency and accuracy still persist, and the current knowledge about pathways and networks is still insufficient to allow a systematic approach.

Another completely different path has been opened by the use of sequence features, (e.g., sequence length, potential phosphorylation sites, predicted TM segments), for the prediction of protein functional class.[31] It is conceivable that in the future these, and other, alternative approaches will be important complements for research in protein functions. However, homology-based function prediction still plays a central role in Molecular Biology.

The Space of Sequences and the Annotation of Function

Our system works by first mapping the sequence space in groups of orthologous sequences. The results of the clustering depend strongly on the quality of the sequence space map built. This, in turn, depends on the parameters used for retrieving the sequences and measuring the distances between them. We have shown previously[23] that this clustering strategy works appropriately for the identification of orthologous groups of sequences from sets of paralogues (families and subfamilies). Compared to other approaches for protein classification, this one has the advantage of being resistant to domain problems because sequence searches are made with respect to a query protein and only the aligned subsequences are used to search the space of sequences. Additionally, it does not require working in the context of complete genomes, as in the case of COGs database.[29]

The procedure has been validated in real biological problems, such as the annotation of the *Buchnera aphidicola*[25] genome described here.

Systems able to make solid predictions of function based on sequence information are important in the context of the annotation of genomes, even if a number of difficulties remain to be solved. The definition of function has a very important subjective component: the same protein will be described in different terms by different scientists. The approach we have followed tries to transfer the most representative description, the consensus definition, which also reduces the pernicious propagation of annotation errors. Domain shuffling events observable at many proteins make function transference based on homology dangerous. An analysis of domain structure is important but, as properly expressed by Attwood,[32] the complexity of biological systems (such as complete proteins) poses important problems for computational approaches because the

properties of a system can be explained by but not deduced from its components (such as protein domains). In this work, we did not analyze the individual components but rather, as a partial solution, assessed the extent to which the similarity covers the entire length of the implied sequences.

ACKNOWLEDGMENTS

We acknowledge the suggestions of O. Olmea for the application of the clustering strategies, and the graph-based representation of the recursive search results. We are grateful to members of the Protein Design Group for interesting discussions and continuous support. Our work has benefited from the interesting ideas on the Ncut algorithm as described in G. Yona's PhD work,[33] and from use of the MESCHACH numerical library, made public by D. E. Stewart and Z. Leyk. We thank Ian Korf for the BPlite BLAST parser. This work has in part been supported by a grant from the Spanish Ministry of Science and Technology (CICYT), and by a fellowship from Madrid's local government.

REFERENCES

1. Altschul SF, Gish W. Local alignment statistics. Methods Enzymol 1996;266:460–480.
2. Smith TF, Zhang X. The challenges of genome sequence annotation or "The devil is in the details". Nature Biotechnol 1997;15: 1222–1223.
3. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. Predicting function: from genes to genomes and back. J Mol Biol 1998;283:707–725.
4. Bork P, Koonin EV. Predicting functions from protein sequences: where are the bottlenecks?. Nature Genet 1998;18:313–318.
5. Doerks T, Bairoch A, Bork P. Protein annotation: detective work for function prediction. Trends Genet 1998;14:248–250.
6. Andrade MA. Position-specific annotation of protein function based on multiple homologs. ISMB 99 1999;7:28–33.
7. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001;307:1113–1143.
8. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
9. Brenner SE. Errors in genome annotation. Trends Genet 1999;15: 132–133.
10. Devos D, Valencia A. Intrinsic errors in genome annotation. Trends Genet 2001;17:429–31.
11. Devos D, Valencia A. Practical limits of function prediction. Proteins 2000;41:98–107.
12. Blaschke C, Hirschman L, Valencia A. Information extraction in molecular biology. Brief Bioinform 2002;3:154–165.
13. Andrade MA, Brown NP, Leroy C, Hoersch S, Daruvar A.De, Reich C, Franchini A, Tamames J, Valencia A, Ouzounis C, Sander C. Automated genome sequence analysis and annotation. Bioinformatics 1999;15:391–412.
14. Moller S, Leser U, Fleischmann W, Apweiler R. EDITtoTrEMBL: a distributed approach to high-quality automated protein sequence annotation. Bioinformatics 1999;15:219–227.
15. Fleischmann W, Moller S, Gateau A, Apweiler R. A novel method for automatic functional annotation of proteins. Bioinformatics 1999;15:228–233.
16. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 2002;3: 265–274.
17. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
18. Reich J, Mitchell A, Goble C, Attwood T. Toward more intelligent annotation tools: a prototype. IEEE Intell Syst 2001;16:42–51.
19. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein database. Bioinformatics 2001;17:282–283.
20. Park J, Teichmann S, Hubbard T, Chothia C. Intermediate sequences increase the detection of homology between sequences. J Mol Biol 1997;273:349–354.
21. Gerstein M. Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. Bioinformatics 1998;14:707–714.
22. Shi J, Malik J. Normalized cuts and image segmentation. Proc IEEE Conf Comp Vision Pattern Recog 1997;731–737.
23. Abascal F, Valencia A. Clustering of proximal sequence space for the identification of protein families. Bioinformatics 2002;18:908–921.
24. Wu Z, Leahy R. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation. PAMI 1993;11:1101–1113.
25. van Ham RCHJ, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernández JM, Jiménez L, Postigo M, Silva FJ, Tamames J, Viguera E, Latorre A, Valencia A, Morán F, Moya A. Reductive genome evolution in buchnera aphidicola. Proc Natl Acad Sci USA 2003;100:581–586.
26. Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SG. 50 million years of genomic stasis in endosymbiotic bacteria. Science 2002; 296:2376–2379.
27. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nature Genet 2000;25:25–29.
28. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 2002;12:368–373.
29. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997;278:631–636.
30. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. Nature 1999;402:83–86.
31. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt HH, Rapacki K, Workman C, Andersen CA, Knudsen S, Krogh A, Valencia A, Brunak S. Prediction of human protein function from post-translational modifications and localization features. J Mol Biol 2002;319:1257–1265.
32. Attwood TK. Genomics. The Babel of bioinformatics. Science 2000;290:471–473.
33. Yona G. Methods for global organization of the protein sequence space. PhD Thesis, Hebrew University. 1999.