Protein Engineering vol.15 no.3 pp.169–183, 2002

Knowledge-based selection of targets for structural genomics

Dmitrij Frishman

Institute for Bioinformatics, GSF—National Research Center for Environment and Health, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany

E-mail: d.frishman@gsf.de

The problem of rational target selection for protein structure determination in structural genomics projects on microbes is addressed. A flexible computational procedure is described that directly incorporates the whole body of annotation available in the PEDANT genome database into the sequence clustering and selection process in order to identify proteins that are likely to possess currently unknown structural domains. Filtering out gene products based on predicted structural features, such as known three-dimensional structures and transmembrane regions, allows one to reduce the complexity of neighbor relationships between sequences and all but eliminates the need for further partitioning of single-linkage clusters into disjoint protein groups corresponding to homologous families. The results of a large-scale computation experiment in which exemplary target selection for 32 prokaryotic genomes was conducted are presented.

Keywords: fold recognition/genome analysis/
sequence clustering/structural genomics

Introduction

Experimental elucidation of a three-dimensional structure for each known protein sequence will hardly ever be possible. Although both DNA sequencing and protein structure determination have become high-throughput technologies, it is still four to five orders of magnitude more expensive to characterize structurally one amino acid than to sequence the three DNA bases coding for it and only for roughly 13 000 out of 500 000 distinctly different protein sequences currently available has a three-dimensional structure been solved. Fortunately, proteins typically come in families and the number of possible folding patterns is limited. Therefore, to achieve a satisfactory structural characterization of the current population of protein sequences, it would be sufficient to solve one representative structure of each type—the structural knowledge about a particular protein domain occurring in a family member can then be extrapolated to other sequences using homology modeling, provided that the degree of similarity is sufficiently high. Several concerted high-speed structure generation projects have been initiated, with the ultimate goal of solving up to 10 000 new structures within the next 5 years and as a result elucidate a representative of each protein fold existing in nature (Sali, 1998). The availability of completely sequenced genomes plays a key role in these efforts. The joint product of genomics and structural biology, known as structural genomics, turned out to be productive and mutually beneficial: while genome data allow for more efficient exploration of the protein structure space (Gerstein, 1998; Frishman and Mewes, 1999; Wolf *et al.*, 2000), prediction and experimental determination of protein structures are crucial for improving functional inferences from genomes (Milburn *et al.*, 1998; Hegyi and Gerstein, 1999; Skolnick and Fetrow, 2000).

One particular aspect of structural genomics involves the systematic structural exploration of complete proteomes. Each genome of a free-living organism codes for a complete set of functions and hence corresponding protein structures necessary to support cellular life. Statistically, structural tendencies in complete genomes, such as the fraction of residues in α-helical and β-sheet conformations, are well conserved between different species, but differ significantly from the observed distribution in the current collection of known protein structures (Frishman and Mewes, 1997a). This observation led us to suggest that determining the complete set of structures encoded in a small model organism would be of great value for structural biology and would have the potential to provide us quickly with a more objective view of the diversity of protein folds. Structural knowledge can also help to decipher the function of the majority of proteins in each genome that cannot be characterized though the application of standard bionformatics approaches, such as similarity searches (Kim, 2000). In particular, genome-wide structure determination is the most direct way to address the problem of genomic ‘ORFans’, i.e. proteins without known function occurring in only one organism (Fisher and Eisenberg, 1999). The benefits of structural genomics on microbes are especially evident with pathogens and also microorganisms adapted to extreme environments because of the immediate relevance for medicine and biotechnology, respectively (Terwilliger *et al.*, 1998). Efforts to obtain complete structural complements of several microbial species are now under way [*Mycobacterium tuberculosis* (http://www.doe-mbi.ucla.edu/TB/index.html), *Pseudomonas aerophilum* (Mallick *et al.*, 2000), *Haemophilus influenzae* (http://s2f.umbi.umd.edu), *Methanococcus jannaschii* (http://sb3.lbl.gov/genomics/proteinlist.html), *Methanobacterium thermoautotropicum* (http://nmr.oci.utoronto.ca/arrowsmith/proteomics/index.html)]. In Japan, the Structurome Project (Yokoyama *et al.*, 2000) pursues the determination of all structures from the thermophilic eubacterium *Thermus thermophilis* (http://www.rsgi.riken.go.jp/). This genome was selected for a large-scale protein structure study because of its compactness, thermostability, presumed ease of crystallization and the availability of genetic tools for further functional essays. Projects of this type are already beginning to bear fruit. For example, Hwang *et al.* assigned a function to a previously uncharacterized gene product of *M.jannaschii* by means of the crystallographic analysis of its three-dimensional structure (Hwang *et al.*, 1999).

At a given state of the technology to determine structures, selection of the most economical set of targets is a major cost-saving factor in any experimental structural genomics project. The principal requirement of any such target list is that it must reveal the minimal collection of gene products that possess all

© Oxford University Press

structural domains with yet unknown folds in the entire data set under study. Given the high abundance of duplication modules, both on the level of whole genes or parts of genes intrinsic to all complete genomes, a crucial step in creating the list of putative structural targets involves grouping together proteins sharing similar sequence segments. This is typically achieved through single-linkage clustering of amino acid sequences based on pairwise similarity comparisons. Owing to the well known phenomenon of domain chaining, totally unrelated protein sequences may end up in the same cluster. Sophisticated approaches have been developed to partition single-linkage clusters further into groups of proteins that are guaranteed to share sequence similarity (Sonnhammer and Kahn, 1994; Koonin *et al.*, 1996; Park and Teichmann, 1998; Matsuda *et al.*, 1999; Yona *et al.*, 1999; Enright and Ouzounis, 2000). However, in many cases joint mosaic occurrence of multiple conserved protein modules (Bork *et al.*, 1997) gives rise to very large sequence groups with a complex structure of inter-sequence similarity relationships. Partitioning the corresponding single-linkage clusters into single domain clusters may represent a significant algorithmic challenge.

Table I. Genome sequences considered in this study

Genome	Domain[a]	Number of ORFs
Aquifex aeolicus	B	1522
Archaeoglobus fulgidus	A	2407
Aeropyrum pernix	A	2694
Borrelia burgdorferi	B	850
Bacillus subtilis	B	4099
Campylobacter jejuni	B	1731
Chlamydia pneumoniae CWL029	B	1052
Chlamydia pneumoniae AR39	B	997
Chlamydia trachomatis serovar D	B	894
Chlamydia trachomatis MoPn	B	818
Deinococcus radiodurans	B	3101
Escherichia coli	B	4277
Haemophilus influenzae	B	1709
Helicobacter pylori	B	1553
Helicobacter pylori J99	B	1491
Mycoplasma genitalium	B	480
Methanococcus jannaschii	A	1735
Mycoplasma pneumoniae	B	677
Methanobacterium thermoautotrophicum	A	1869
Mycobacterium tuberculosis	B	3924
Neisseria meningitidis MC58	B	1989
Pseudomonas aeruginosa	B	5565
Pyrococcus abyssi	A	1765
Pyrococcus horikoshii	A	2064
Rickettsia prowazekii	B	834
Synechocystis sp.	B	3169
Thermoplasma acidophilum	A	1509
Thermotoga maritima	B	1846
Treponema pallidum	B	1031
Ureaplasma urealyticum	B	611
Vibrio cholerae	B	1092
Xytella fastidiosa	B	2765

See http://pedant.gsf.de/credits.html for a list of the Web links to the respective sequencing centers.
[a]A, Archaea; B, Eubacteria.

It should be noted that the sequence clustering tools mentioned above were developed with the purpose of studying the family relationships between proteins for better functional, structural and evolutionary inferences. While this information is also invaluable in the context of target selection for structural genomics, the immediate technical objective here is much more limited: we want to exclude protein domains that either do not belong to our targeted class (e.g. transmembrane proteins if we are interested in soluble proteins) or already have been structurally characterized. In this paper, we argue that the computational complexity of the target selection process can be significantly reduced if the knowledge about predicted structural features and other relevant protein properties is directly incorporated into the clustering procedure. It is sufficient to perform the simple step of initial single-linkage clustering. After the application of a number of filtering criteria, many of these clusters will be excluded from consideration because all sequences constituting them have been discarded. In some other cases, single-linkage clusters will be reduced to just one candidate sequence. Finally, the remaining clusters will include sequences all of which are potential structural targets. The complicated procedure of resolving the domain structure of single-linkage clusters thus becomes obsolete. An important prerequisite of this approach is the availability of a comprehensive annotated database of completely sequenced genomes.

Materials and methods

Genome sequences and annotation

In this work we considered completely sequenced genomes of 25 eubacterial and seven archaebacterial species (Table I). Exhaustive automatic annotation of these genomes was conducted using the PEDANT genome analysis suite (Frishman and Mewes, 1997b; Frishman *et al.*, 2001) and can be accessed through the PEDANT genome database (http://pedant.gsf.de).

The main distinctive feature of the PEDANT system is its ability to assign proteins to automatically derived structural and functional categories. The categorization system is multidimensional in that each sequence can be assigned to many different categories and each category can contain any number of gene products. The main vehicle for similarity searches against the full non-redundant protein sequence database and a number of specialized datasets (e.g. functional categories) is the PSI-BLAST algorithm developed at the National Center for Biotechnology Information, Bethesda, MD (Altschul *et al.*, 1997). In addition, detection of various sequence motifs, extraction of relevant keywords, enzyme classification and superfamily information are performed.

Structural categorization of gene products involves a highly sensitive comparison of each gene product with the SCOP database of known structural domains (Brenner *et al.*, 2000;

Fig. 1. Flow chart of the target selection algorithm, exemplified using the *E.coli* genome. The entire protein complement of *E.coli*, comprising 4277 gene products, is subjected to single-linkage clustering and is split into 2235 clustered sequences and 2042 singlets (sequences not having any paralogs in the genome). Both subsets are filtered to exclude transmembrane proteins and those proteins which are completely structurally characterized. The resulting singlets are declared structural targets. The remaining 701 sequences are subjected to two stages of iterative re-clustering. In the first stage, simple domain problems, such as cannibalization of a short domain by a longer one, are resolved. In the second stage, more complex situations involving proteins with domain similarities to protein of known structure are treated. At each iteration, redundant sequences are excluded from further consideration and clusters reduced to just one sequence become singlets and are added to the target pool. Sequences remaining in the single-linkage clusters after the application of the algorithm are also declared structural targets since they are guaranteed to possess at least one unique and sufficiently long domain not covered by structural information.

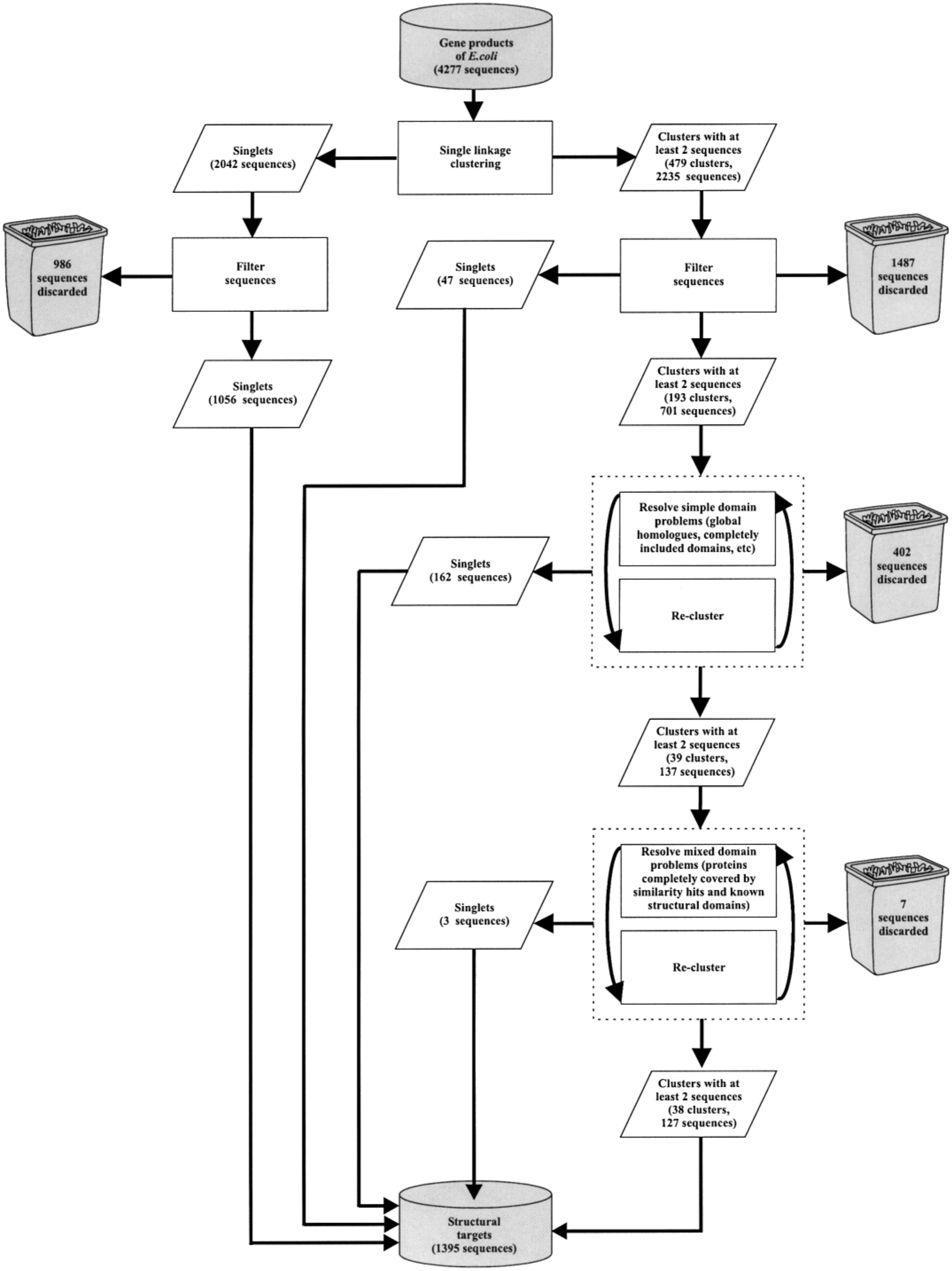
Gene products
of *E.coli*
(4277 sequences)
Single linkage
clustering
Singlets
(2042 sequences)
Clusters with at
least 2 sequences
(479 clusters,
2235 sequences)
986
sequences
discarded
Filter
sequences
Singlets
(47 sequences)
Filter
sequences
1487
sequences
discarded
Singlets
(1056 sequences)
Clusters with at
least 2 sequences
(193 clusters,
701 sequences)
Resolve simple domain
problems (global
homologues, completely
included domains, etc)
Singlets
(162 sequences)
Re-cluster
402
sequences
discarded
Clusters with at
least 2 sequences
(39 clusters,
137 sequences)
Resolve mixed domain
problems (proteins
completely covered by
similarity hits and known
structural domains)
Singlets
(3 sequences)
Re-cluster
7
sequences
discarded
Clusters with at
least 2 sequences
(38 clusters,
127 sequences)
Structural
targets
(1395 sequences)

Fig. 1.

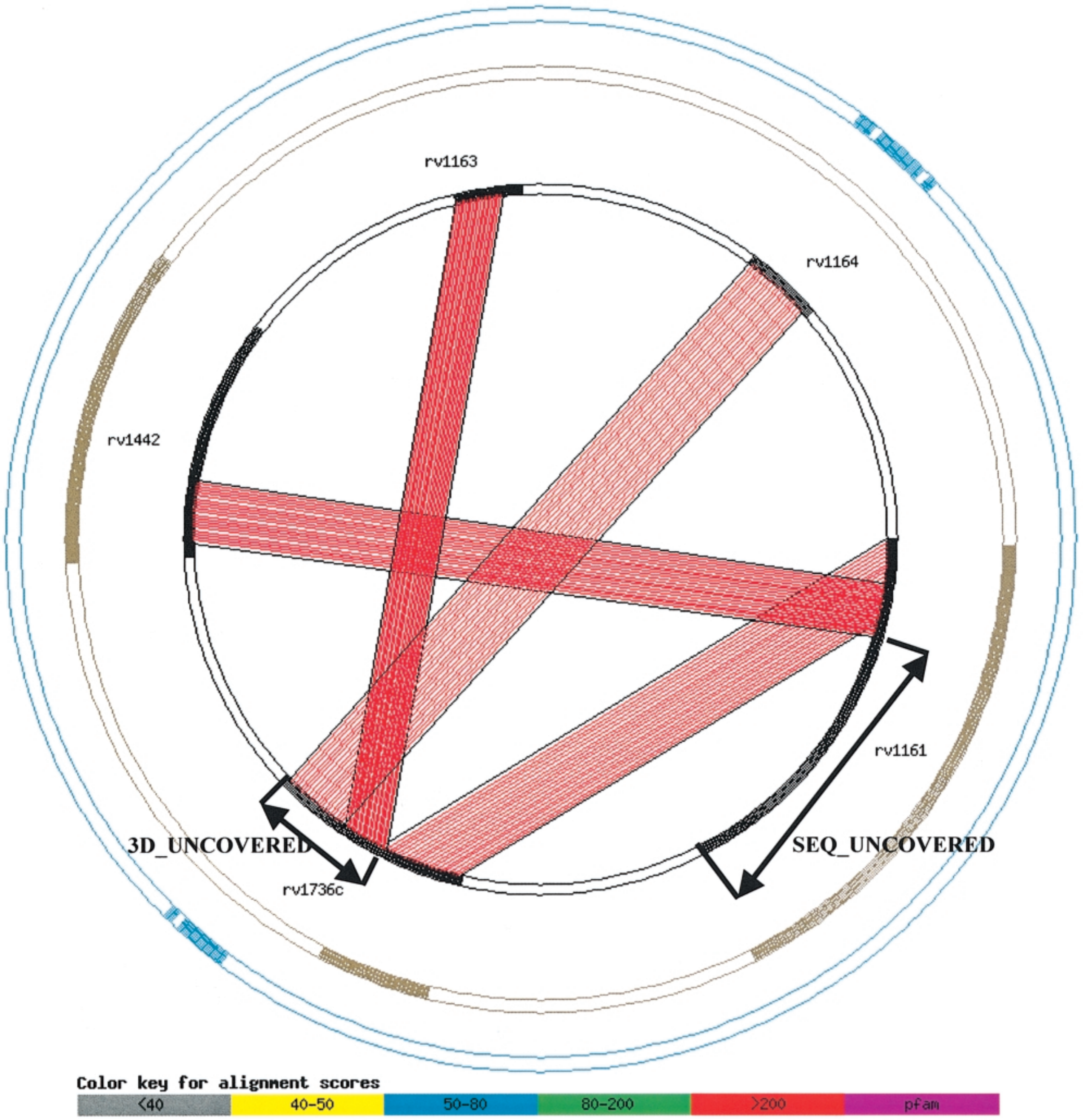

Fig. 2. Circular representation of a single-linkage cluster (circlegram). On the inner circle, five *M.tuberculosis* gene products (*rv1164*, *rv1161*, *rv1736*, *rv1442* and *rv1163*) are shown as black sectors. The N- to C-terminal direction is clockwise. Sequence regions aligned by BLAST after an all-against-all comparison of *M.tuberculosis* proteins are joined by stripes colored according to the BLAST alignment score. On the next (middle) circle, shown in brown, IMPALA hits in the database of protein sequences with known three-dimensional coordinates are mapped. The outer circle, shown in blue, indicates the location of predicted transmembrane regions. Positions of other functional and structural features, such as SCOP domains, protein motifs, low-complexity regions, etc., can be shown on further concentric circles, one for each feature.

Lo Conte *et al.*, 2000) and the sequences of proteins with known three-dimensional structure (Berman *et al.*, 2000) using the novel IMPALA software (Schaffer *et al.*, 1999). This program allows one to compare a query protein sequence with a collection of position-specific scoring matrices generated by BLAST and is thus perfectly suitable for similarity-based fold recognition (Wolf *et al.*, 1999). Our current approach to genomic fold recognition involves the following steps: (i) create a complete non-redundant protein sequence database, (ii) run a PSI-BLAST search with 10 iterations with each SCOP domain or PDB sequence against the non-redundant protein sequence database and save the resulting profiles, (iii) construct a SCOP or PDB profile library using the IMPALA software suite and (iv) run an IMPALA search with each genomic sequence against the SCOP or PDB library. Additionally, for each genomic sequence a number of structural features are predicted, including secondary structure (Frishman and Argos, 1997), low-complexity regions (Wootton and Federhen, 1993), membrane regions (Klein *et al.*, 1985), coiled coils (Lupas *et al.*, 1991) and signal peptides (Nielsen *et al.*, 1997).

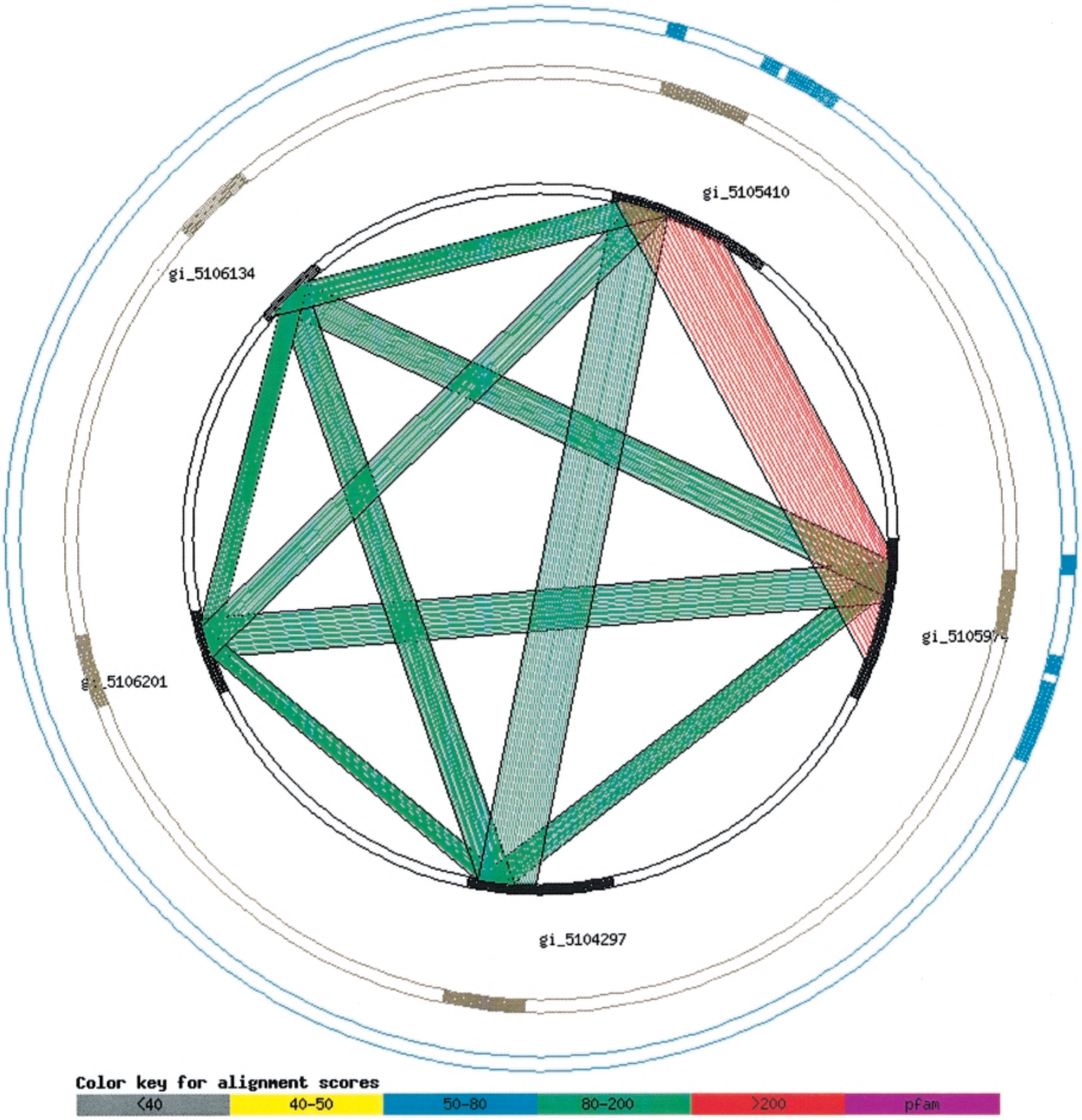

Fig. 3. Single-linkage cluster involving five gene products from *A.pernix*.

Single-linkage clustering of complete genomic protein complements

An all-against all comparison of proteins within each genome was effected using PSI-BLAST, with low-complexity sequence regions masked. Sequences possessing a sufficient degree of similarity in a reciprocal fashion (BLAST similarity score >45 bits) were joined into single-linkage groups. In cases where reciprocal BLAST comparisons produced only one local alignment between two sequences in each direction, this hit was made symmetrical by taking into account only the longer alignment. Optionally, it is also possible to take into account results of sensitive recognition of PFAM domains (Bateman *et al.*, 2000) through HMMER searches (Eddy, 1998). If two or more proteins in a genome display similarity to the same PFAM domain with a significant *E* value (typically 0.001), it may be safely assumed that the corresponding protein sequence spans are similar to each other, even if BLAST fails to recognize such relationships.

The lists of clustered sequences for all identified clusters in 32 completely sequenced genomes are available through the PEDANT Web site (see category sequence clusters).

Algorithm for target selection

The flow chart of our algorithm, which we dub STRUDEL (STRUcture DEtermination Logic), is presented in Figure 1, using the genome of *Escherichia coli* as an example. The analysis of each protein complement begins with single-linkage clustering. As a result, all sequences are partitioned into two sets: singlets, i.e. sequences without homology to other gene products of the genome considered and hence not participating in any cluster and clustered sequences. Both sets are subjected to filtering according to user-specified criteria. Throughout this work we excluded from further consideration predicted transmembrane proteins and sequences with known three-dimensional structure. More specifically, sequences were filtered out if they (i) had more than one predicted transmembrane region or (ii) the maximum length of a sequence

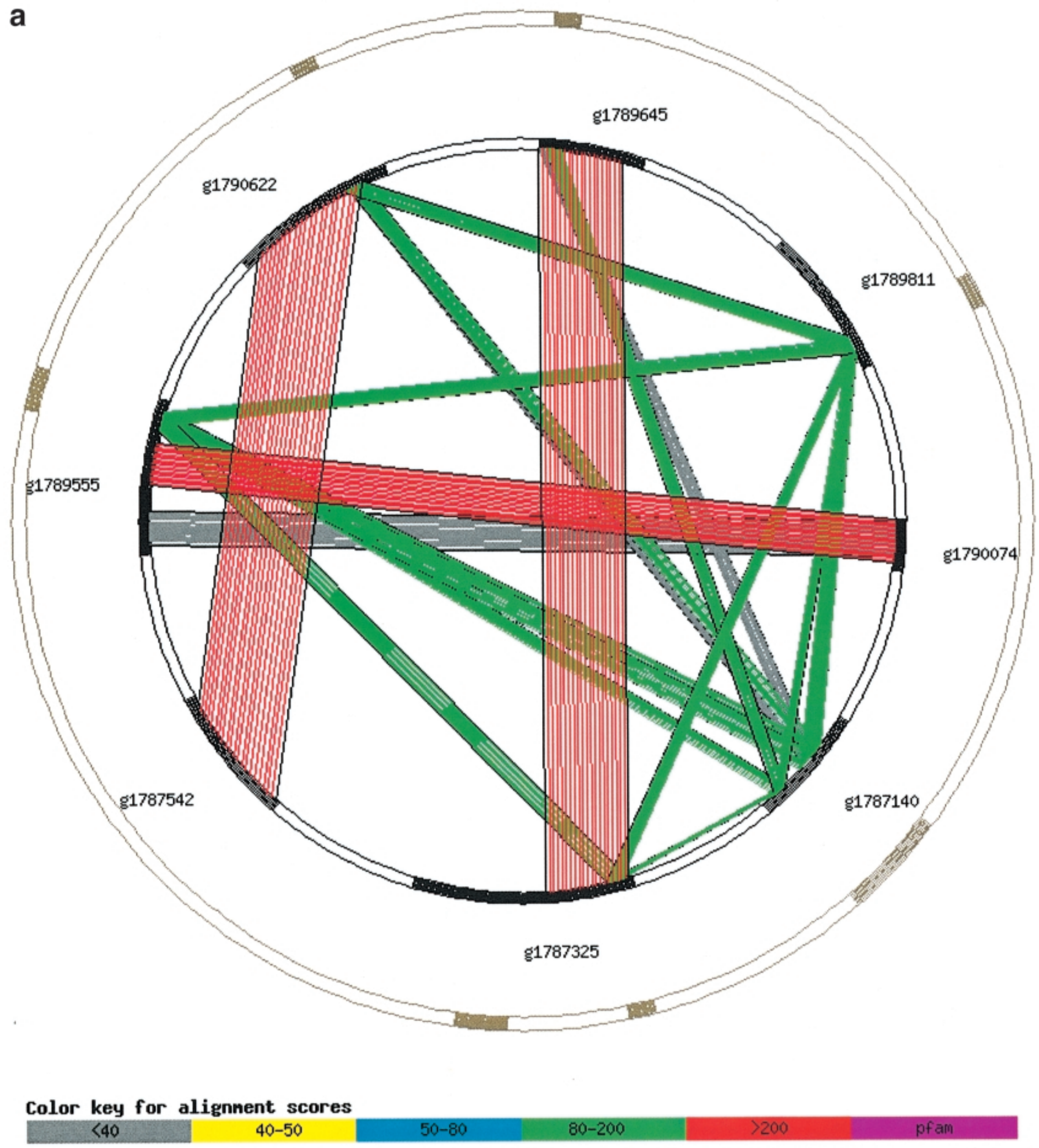

Fig. 4a.

span not covered by IMPALA similarity hits to proteins of known structure was below a certain threshold (denoted 3D_UNCOVERED; Figure 2), reflecting the expected length of a structural domain. Singlets remaining after the filtering are considered structure determination targets. The entire pool of sequences possessing paralogs is re-clustered and a number of newly created singlets are attributed to the structural target set.

In the subsequent steps of the procedure, the structure of sequence alignments among the members of single linkage clusters and the similarity hits to proteins of known structure are analyzed with the goal of excluding redundant information. First, simple domain problems, such as cannibalization of a short domain by a longer domain and completely duplicate, globally similar sequences, are handled. The parameter SEQ_UNCOVERED determines the maximum allowed length of the contiguous sequence span not involved in local alignments with other proteins in the cluster. Next, more complex situations involving the mapping of known three-dimensional domains on the alignments between clustered sequences are resolved. Sequences that have less than SEQ_UNCOVERED amino acid residues not covered by either three-dimensional hits or alignments with other sequences are excluded because some of their domains already have structural information while the remaining sequence portions are completely contained in the other cluster members. The iterative process of discarding superfluous sequences from sequence clusters, one at a time, and re-clustering the remaining sequences continues until convergence, i.e. either until no more sequences can be discarded using the criteria currently applied or the cluster is reduced to just one sequence. In the latter case the resulting singlet is declared the structural target. The resulting clusters include polypeptide chains that have at least 3D_UNCOVERED residues without structural information and at least SEQ_UNCOVERED residues not covered by either structural or similarity hits. As seen in Figure 1, the number of the single-linkage clusters left decreases significantly after each round of exclusion and re-clustering.

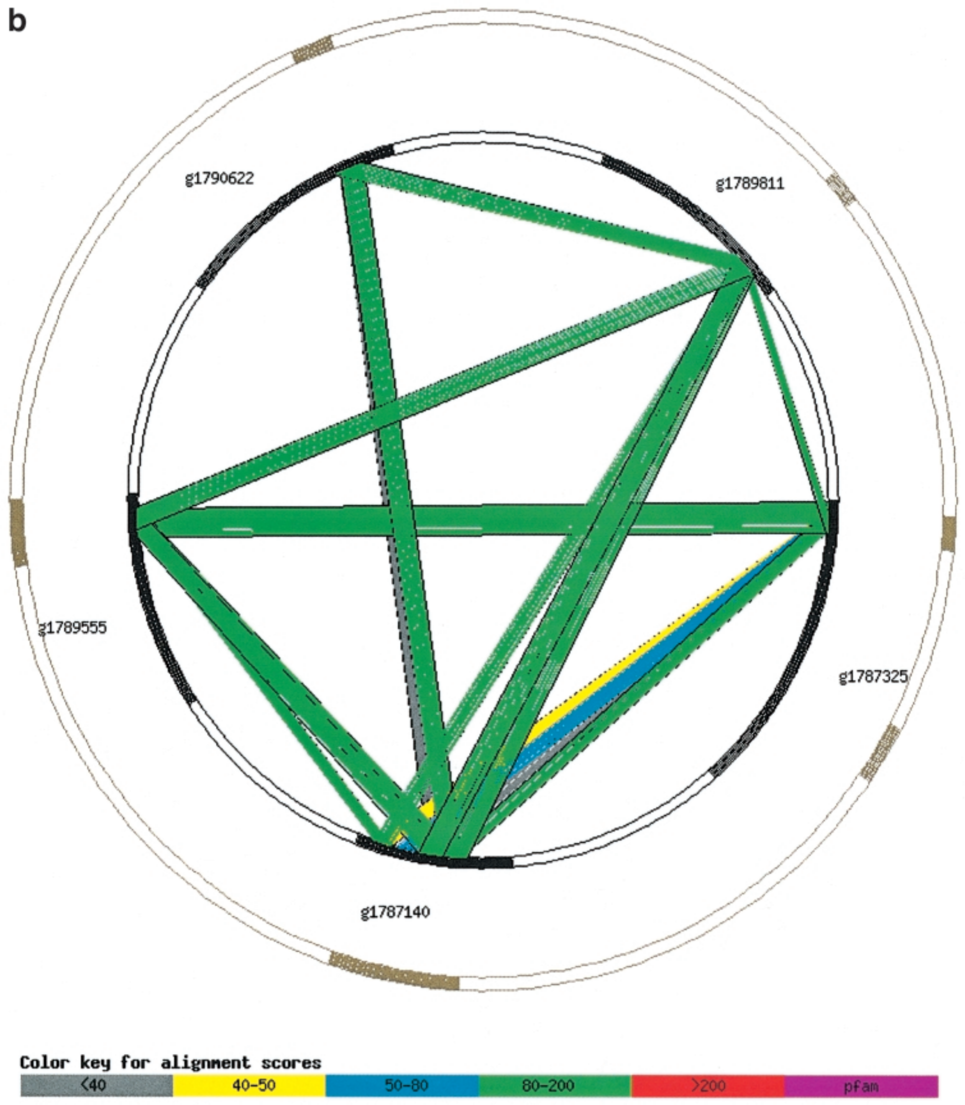

Fig. 4. *Continued.*

The default values of SEQ_UNCOVERED and 3D_UNCOVERED used in this work are both 100 amino acid residues, corresponding to the expected favorable size of a structural domain (Xu and Nussinov, 1998). The influence of these parameters on the results of target selection is detailed below.

The particular succession of steps shown in Figure 1 is not mandatory and is mainly dictated by simple practical considerations. For example, it would be possible to do the preliminary sequence filtering first and then cluster the remaining proteins. However, clustering requires much more time than filtering and it is much more likely that the user of the system will want to change filtering parameters than clustering parameters. Hence it is sensible to do the clustering of the complete set of proteins only once and save the result in the PEDANT relational database. Subsequent steps of the analysis can be quickly performed according to user-specified filtering conditions.

Graphical representation of single-linkage clusters

Owing to the multi-domain composition of many proteins, single-linkage clusters often include sequences that are totally unrelated to each other. In order to facilitate the analysis of the resulting groups we have implemented a visual representation of single-linkage clusters, further referred to as circlegram (Figure 2). A circlegram may include any number of concentric circles, each for a certain protein feature. In this work, the inner-most circle on such graphics schematically represents polypeptide chains as black sectors, with the N- to C-terminal direction corresponding to the clockwise direction on the circlegram. On the next circle of larger radius, IMPALA similarity hits to proteins of known structure are depicted as brown sectors. Finally, the outermost circle indicates the location of predicted transmembrane regions in blue. Owing to the small scale of the graphics and relatively low resolution, several features, e.g. transmembrane domains, may be lumped into one contiguous sector. BLAST similarity hits between the proteins constituting the cluster are shown as stripes originating from respective black sectors, with boundaries corresponding to the start and end positions of local alignments. The stripes are colored according to the BLAST similarity scores (see the color key in the bottom of each picture). Similarity relationships between proteins based on the presence of PFAM domains (see above) may be additionally shown as stripes of a single color, irrespective of the *E* values of the underlying PFAM search hits. Circlegrams represent a convenient means of

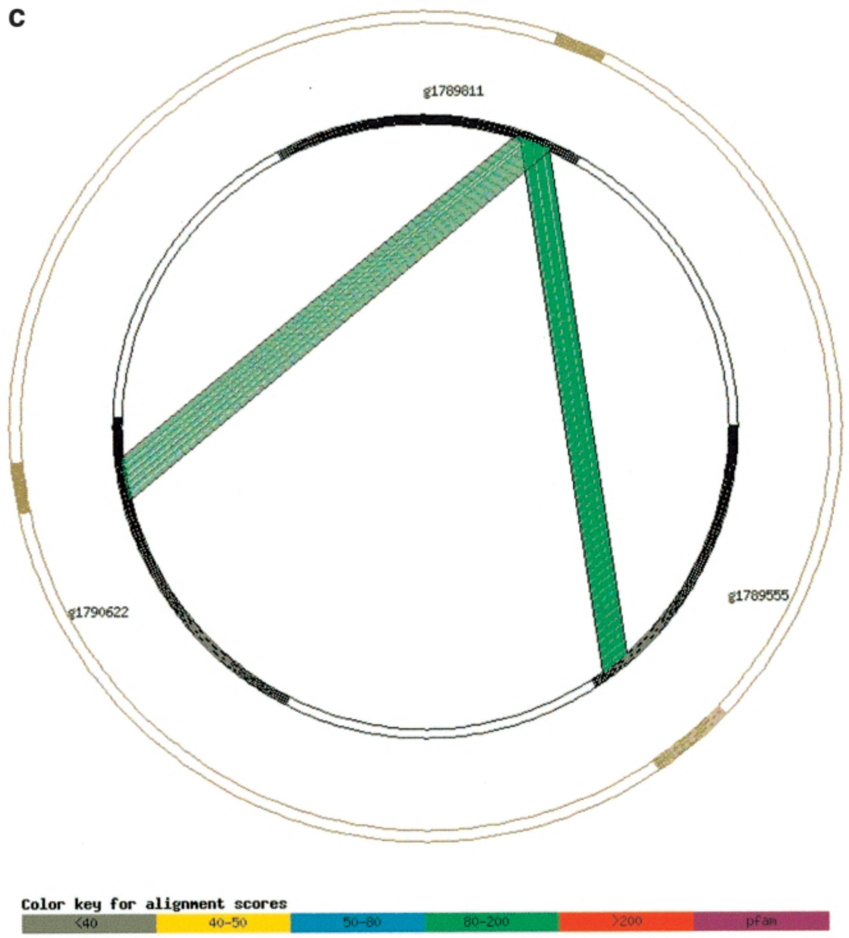

Fig. 4. *Continued.*

displaying any number of sequence-related structural and functional features together with intra-protein similarity relationships. They are similar in spirit to the circular depiction of correlated structural features within one protein sequence developed by Pazos *et al.* (Pazos *et al*., 1997) and complementary to the linear diagrams of domain similarity implemented by Storm and Sonnhammer (Storm and Sonnhammer, 2001).

Results and discussion

Combining sequence clustering with the analysis of protein structural features

The main distinctive feature of our method is the direct incorporation of predicted protein structural features into the clustering procedure. This approach allows to discard a large number of gene products at early stages of the target selection process and radically reduce the complexity of the resulting single-linkage clusters. Figure 3 provides an example of a single-linkage cluster from *Aeropyrum pernix* which is collapsed to a singlet if standard settings described in the Materials and methods are applied. All sequences forming the cluster possess the thioredoxin domain with known tertiary structure. Two genes, gi_5106134 and gi_5106201, code for apparently single domain proteins and are discarded since they are almost completely covered by three-dimensional information. Two further proteins, gi_5105410 and gi_5105974, will be discarded since they have a large membrane-spanning domain on the C-terminus and an additional putative hydrophobic region on the N-terminus. The remaining gene product gi_5104297 is retained as a structural target because in addition to the C-terminal thioredoxin domain it includes a completely uncharacterized soluble domain on the N-terminus. The results for this cluster will be different if the user is interested in shorter domains and sets both 3D_UNCOVERED and SEQ_UNCOVERED to ~75 amino acid residues. Then gi_5106201 also becomes a structural target since its N-terminal portion encompassing approximately the first 80–90 amino acids displays no similarity to any other known protein.

The *E.coli* cluster shown in Figure 4a involves the 30S ribosomal protein S1 (g1787140) (Kimura *et al.*, 1982) and a number of other RNA-associated proteins. The S1 protein contains six copies of the S1 RNA binding domain (of which IMPALA recognizes only four) with the three-dimensional structure solved by Bycroft *et al.* (Bycroft *et al.*, 1997). Single occurrences of this domain are also detected in five other proteins in this cluster, g1790622, g1789645, g1789811, g1787325 and g1789555. In the latter, another DNA-binding domain, the KH module (Siomi *et al.*, 1993), with known structure, is also present immediately adjacent to the S1 domain. In addition, the ribonuclease E protein (g1787325) has a very weak IMPALA hit (score = 45 bits, *E* value = 0.004) to another protein with known structure, the NCD kinesin motor protein from *Drosophila melanogaster* shown in the central part of the protein; this similarity is certainly

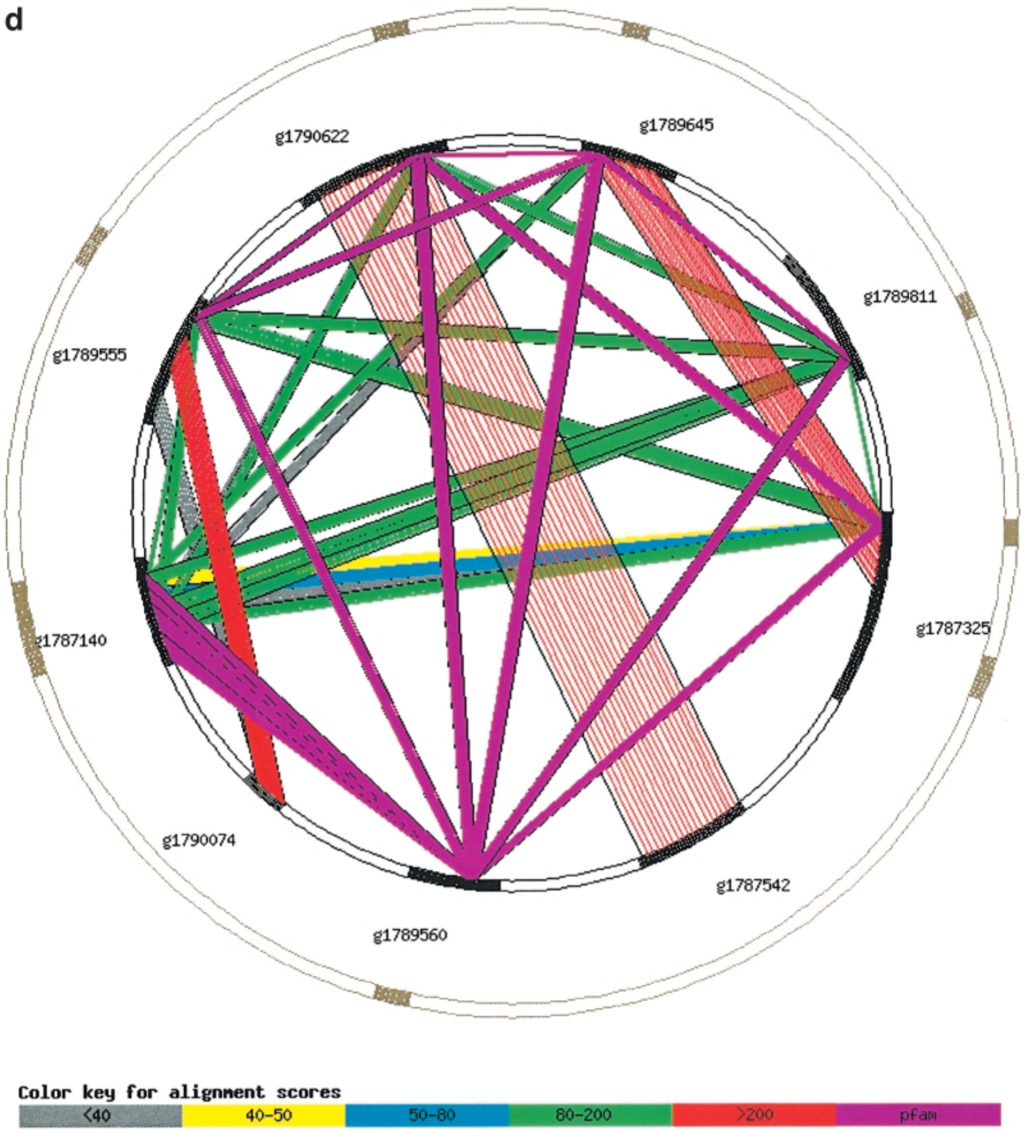

Fig. 4. *Continued.*

spurious. Two further proteins, g1787542 and g1790074, do not contain any domains with known structure, but share a domain sequence similarity with g1790622 and g1789555, respectively. Using the default parameters, the STRUDEL algorithm rejects g1787542, g1789645 and g1790074 because they are completely contained in other proteins of this group. As a result, five proteins with partially known structures form the final cluster (Figure 4b) and all of them are declared structural targets. In fact, however, g178140 must have been rejected because it is completely covered by the six copies of the S1 domain. The problem is that, as mentioned above, only four of them are detected by the IMPALA search so that over 150 amino acid residues of this protein end up not having structural information. Increasing the 3D_UNCOVERED parameter to 160 residues leads to g178140 being discarded, but at the same time g1787325 is also discarded because in this case two structural hits, the correct one to the S1 domain and the incorrect one to the NCD protein, account for a sufficiently large fraction of its polypeptide chain (Figure 4c). Thus, using 3D_UNCOVERED = 100 and 3D_UNCOVERED = 160 in this example leads to over- and under-prediction of potential structural targets, respectively. The correct answer can only be achieved by imposing a stricter similarity threshold for IMPALA searches (score = 45 bits) in combination with 3D_UNCOVERED = 160. The spurious similarity between g1787325 and the NCD protein will then be below the threshold and g1787325 is recognized as a structural target because only a minor part of its polypeptide chain is covered by the similarity hit to the S1 domain.

The results are different if the PFAM similarity data are considered while building single-linkage clusters (see Materials and methods). One more protein, the L factor protein with the PEDANT id g1789560, is recruited because the PFAM motif corresponding to the S1 RNA binding domain is identified based on HMMER searches (Figure 4d). The relatedness of this protein to other members of the cluster could not be detected through the BLAST all-against-all comparisons. In this case the resulting cluster after the application of STRUDEL encompasses six potential structural targets, one more than without PFAM (Figure 4e). However, the total number of targets derived by the algorithm from this group of proteins remains unchanged. In the case when PFAM hits are not

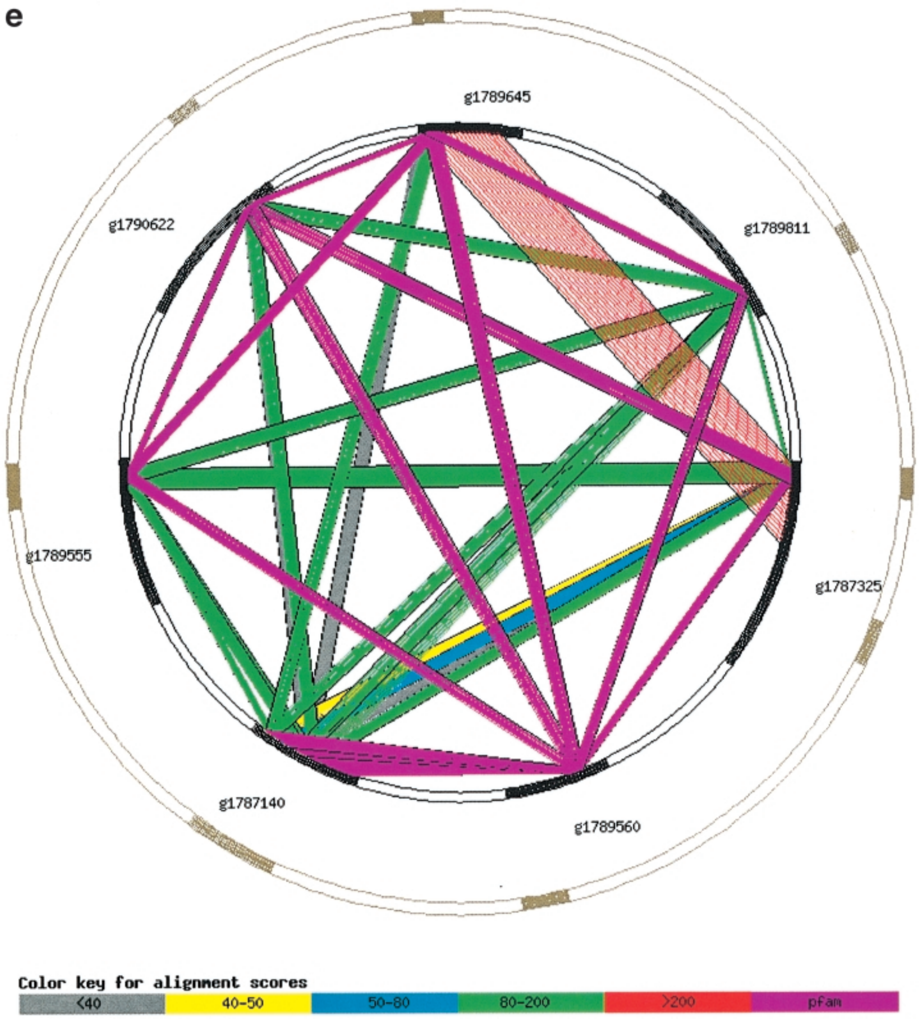

Fig. 4. The *E.coli* cluster involving the 30S ribosomal protein S1 and related proteins. (**a**) Initial single-linkage cluster obtained without consideration of PFAM hits; (**b**) resulting single-linkage cluster after the application of STRUDEL with default parameters; (**c**) resulting single-linkage cluster after the application of STRUDEL with 3D_UNCOVERED = 160 amino acid residues; (**d**) initial single-linkage cluster obtained taking into consideration PFAM hits; (**e**) resulting single-linkage cluster after the application of STRUDEL.

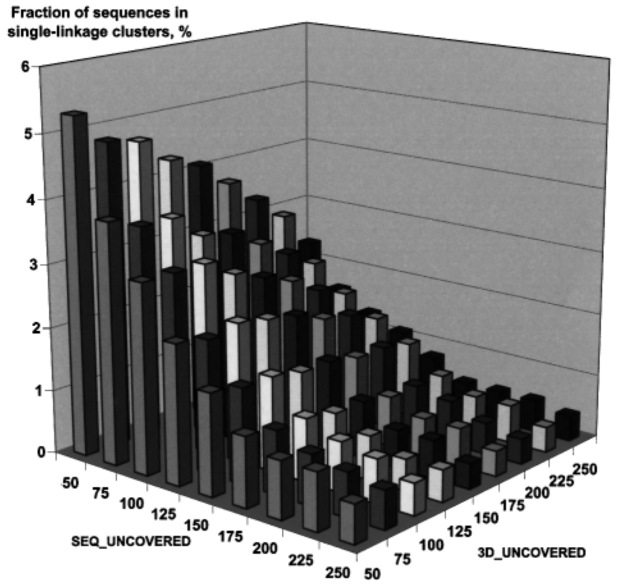

Fig. 5. Dependence of the number of sequences remaining in single-linkage clusters after the application of the STRUDEL algorithm (Figure 1) on the parameters SEQ_UNCOVERED and 3D_UNCOVERED.

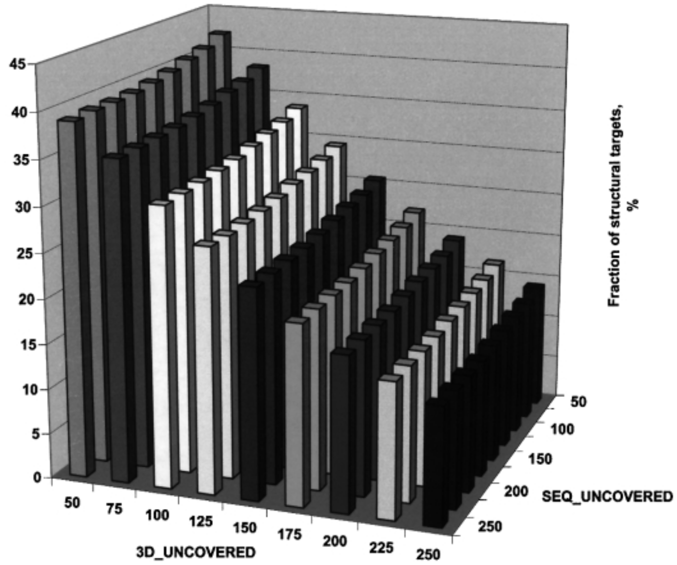

Fig. 6. Dependence of the number of structural targets, among both clustered sequences and singlets, produced by the STRUDEL algorithm on the parameters SEQ_UNCOVERED and 3D_UNCOVERED.

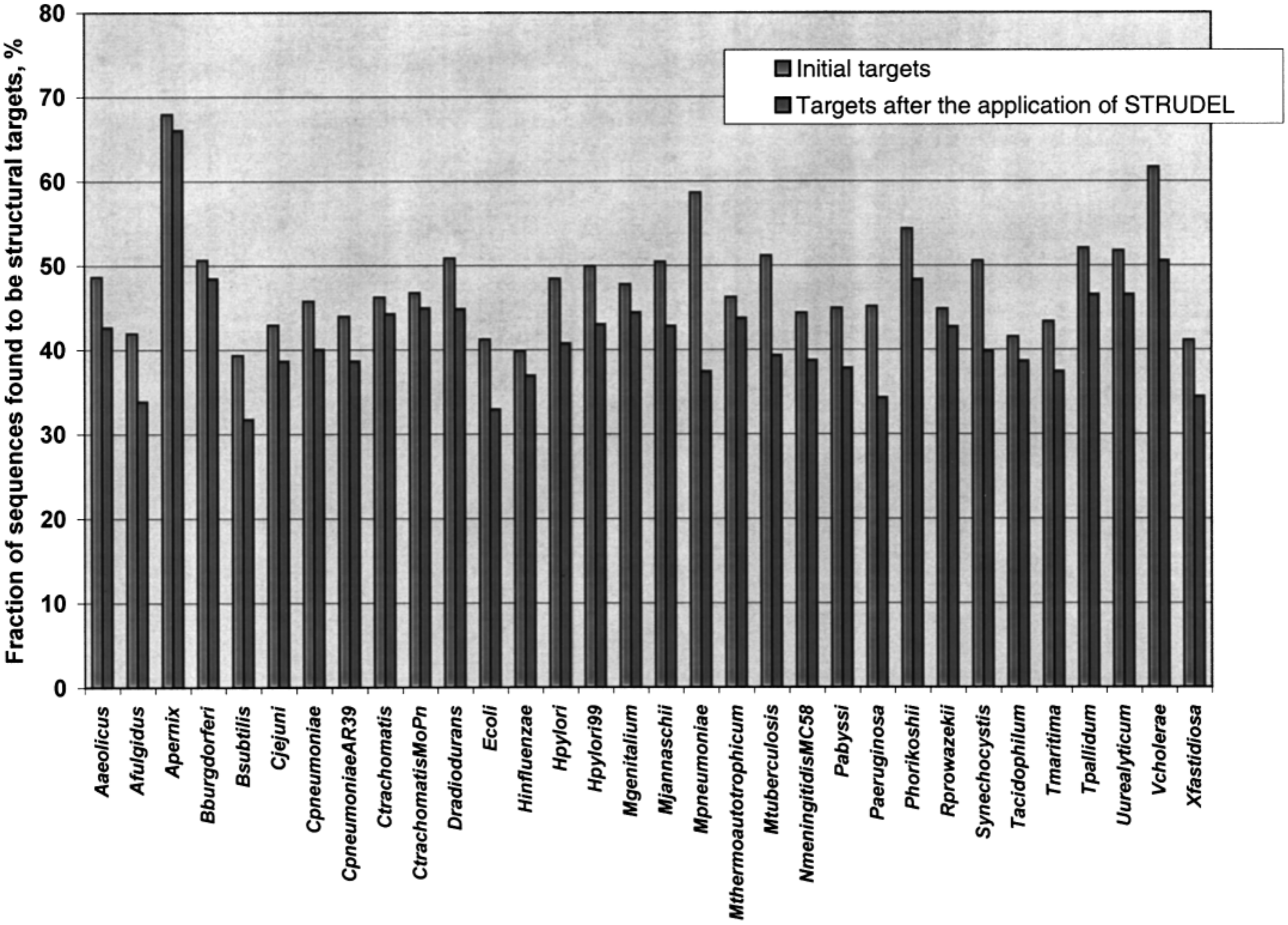

Fig. 7. Percentage of structural targets in 32 complete genomes before and after the application of STRUDEL.

considered g1789560 does not get assigned to any single linkage cluster and ends up being a singlet. Since only a minor part of this protein is covered by the S1 domain with known structure, it is also declared to be a structural target. Thus, in this particular example consideration of PFAM does not have any influence on the number of targets generated, although the better knowledge about the composition of this protein family is certainly helpful for subsequent manual evaluation of the results.

Depending on the objectives of a particular structure determination project, the requirement that the potential structural targets must not have significant transmembrane domains may be relaxed in order to take into account individual, sufficiently long globular domains of membrane-associated proteins. The *M.tuberculosis* sequence cluster centered around the ‘fused nitrate reductase’ *rv1736c* (Figure 2) provides a good illustration of this point. *rv1736c* is the result of re-arrangement and fusion of the α, δ and γ chains of membrane-bound nitrate reductase, encoded by genes *rv1161* and *rv1163* and *rv1164*, respectively (Cole *et al.*, 1998). The soluble α subunit (together with the β subunit) is anchored to the plasma membrane by the γ subunit, while the δ polypeptide is not part of the final enzyme and is presumably important for the stability of the αβ complex prior to its membrane attachment (Moreno-Vivian *et al.*, 1999). *rv1161*, in its turn, shares a weak domain similarity with the biotin sulfoxide reductase *rv1442*. Since the tertiary structure of the entire *rv1161* gene product is known via its homology to the *R.capsulatis* dimethyl sulfoxide reductase (Schneider *et al.*, 1996), the fused protein rv1736 efficiently consists of one globular domain with known structure, one more globular domain with unknown structure, corresponding to *rv1163* and one transmembrane domain on the C-terminus, corresponding to *rv1163*. Consequently, STRUDEL yields *rv1163* as the only target from this cluster. However, if *rv1163* did not exist, the corresponding globular domain in *rv1736c* would have been overlooked because of the presence of the transmembrane domain in the latter. We have implemented an option in the STRUDEL software that allows to consider mixed membrane/soluble proteins as potential structural targets.

Choice of the analysis parameters

The interactive application of STRUDEL allows to use what-if scenarios to explore different outcomes for a given cluster or for the genome as a whole and to optimize the analysis parameters to suit the goal of a particular structure determination project. In this section we demonstrate the global influence of the parameters 3D_UNCOVERED and SEQ_UNCOVERED on the results of the target selection process using the *E.coli* genome as an example.

The total number of single-linkage clusters found in *E.coli* using the reciprocal BLAST score threshold of 45 bits is 479, encompassing 2235 proteins. As seen in Figure 1, after the first round of filtering (removing sequences with completely known three-dimensional information and membrane proteins), 701 sequences, or 16% of the protein complement, remain clustered in 193 clusters. The next stage of the algorithm, elimination of global redundancy between clustered sequences,

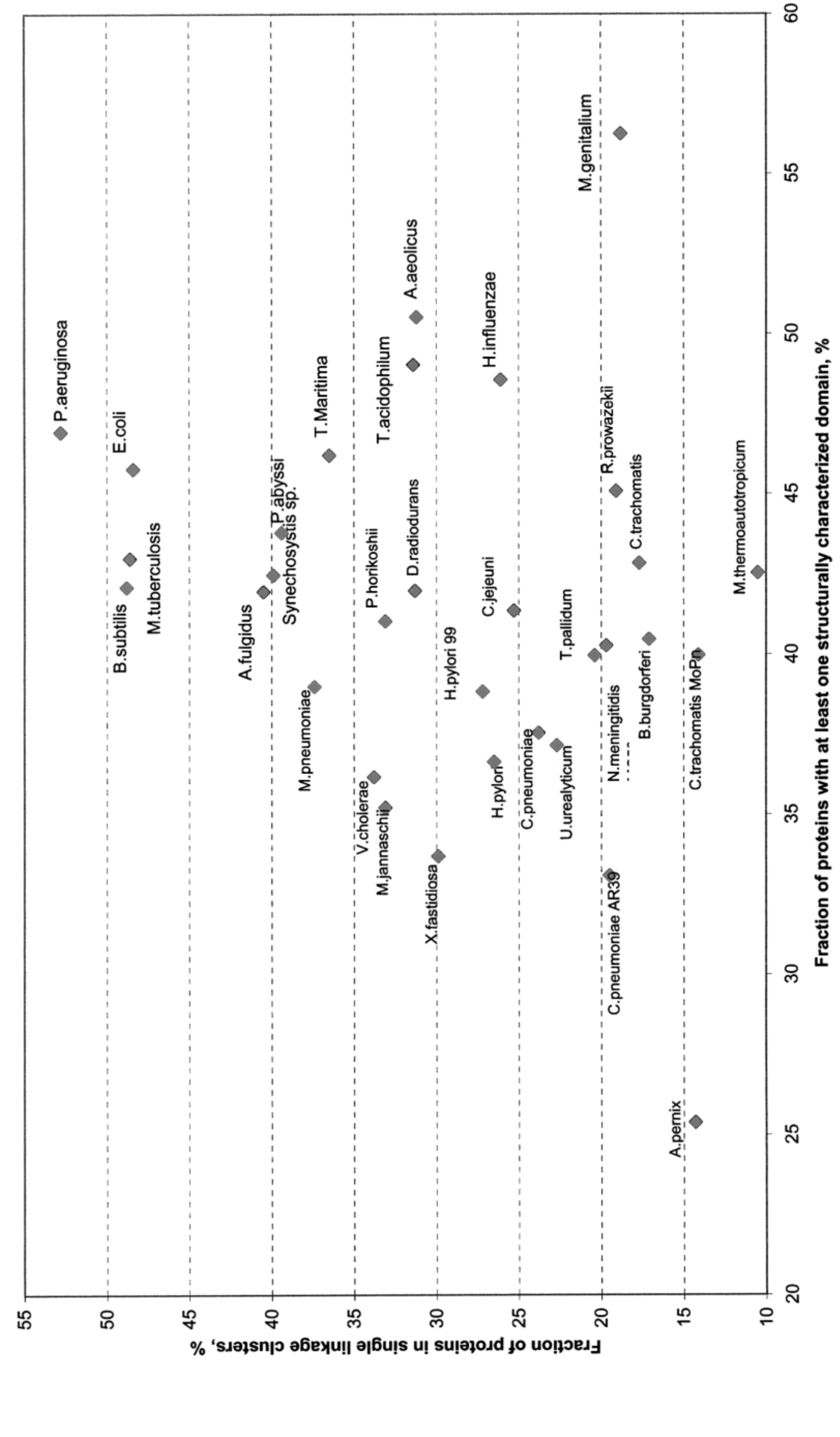

Fig. 8. Comparison of 32 completely sequence genomes in terms of the number of proteins with known structural information versus the percentage of proteins in single-linkage clusters.

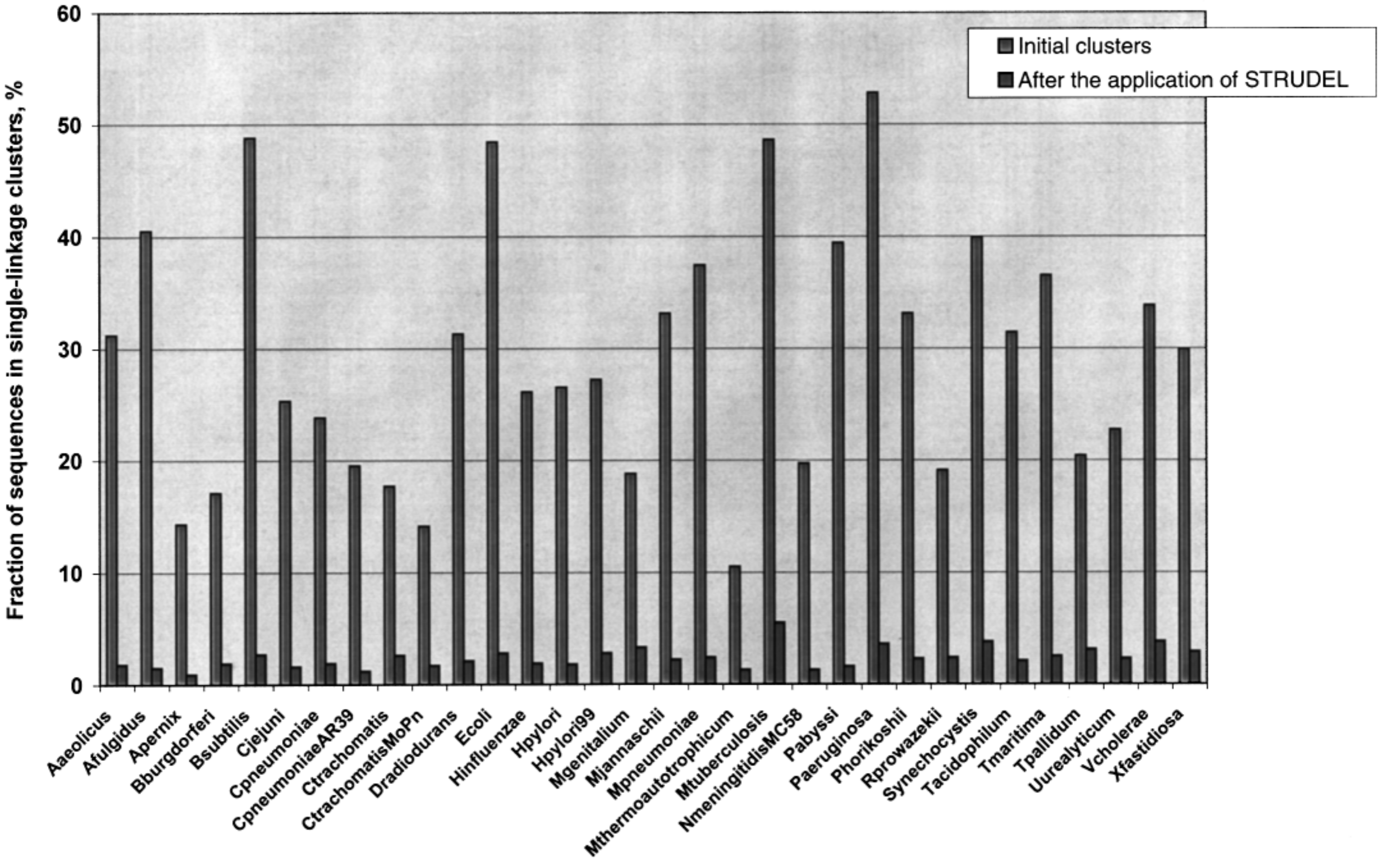

Fig. 9. Percentage of sequences in single-linkage clusters before and after the application of STRUDEL.

results in reducing the number of clustered sequences and clusters to a mere 139 and 39, respectively. Finally, resolving mixed domain problems has a rather insignificant effect, leading to 127 sequences in 38 clusters. Thus, the application of STRUDEL with default parameters (3D_UNCOVERED = 100 and SEQ_UNCOVERED = 100) reduces the fraction of *E.coli* sequences grouped in single-linkage clusters from 52.3% to a mere 3% of the protein complement.

The parameters 3D_UNCOVERED and SEQ_UNCOVERED strongly influence the outcome of the re-clustering procedure (Figure 5). Setting both of them to 50 amino acid residues, for example, results in an increase in the number of clustered sequences to >5% while raising the parameter value to 250 residues eliminates sequence clusters nearly completely (0.4%). Therefore, we conclude that the whole issue of sequence clustering is only of importance if one is interested in structural information on relatively short sequence domains of multi-domain proteins.

The dependence of the total number of structural targets produced by STRUDEL, among both clustered sequences and singlets, on the parameters 3D_UNCOVERED and SEQ_UNCOVERED is shown in Figure 6. It is immediately clear that 3D_UNCOVERED is crucial whereas SEQ_UNCOVERED has a very minor effect on the results. For example, changing SEQ_UNCOVERED from 50 to 250 (with 3D_UNCOVERED = 100) leads to a reduction in the number of targets generated from 1442 to 1327 (which is equivalent to ~3% of the complete gene complement). As shown in the flow chart in Figure 1, using the default parameters, nearly four times more clustered sequences are discarded based on structural filtering criteria (membrane regions, known structural domains) than due to sequence redundancy. Using the default parameters 3D_UNCOVERED = 100 and SEQ_UNCOVERED = 100, 32.6% of the *E.coli* proteins possesses at least one structurally uncharacterized domain. This value varies from ~40% with 3D_UNCOVERED = 50 to as few as 13% with 3D_UNCOVERED = 250. The latter setting is essentially equivalent to focusing only on single-domain proteins or multi-domain proteins in which none of the domains has a known structure.

Genome comparison in terms of the number of structural targets

Figure 7 provides a comparison of the fraction of structural targets, both in singlets and in single-linkage clusters, in all completely sequenced bacterial genomes. On average, 48% of gene products in a genome are globular proteins with at least one structurally uncharacterized domain. After elimination of redundancy on the domain level this figure is lowered to 41.5%. Hence the application of STRUDEL results in the reduction of the number of structural targets by ~7.5%, on average, with respect to the situation where sequence clusters are not taken into account.

The particular values of the number of structural targets for each genome are mostly determined by the interplay of two main factors: the degree of redundancy and the number of known three-dimensional structures identified (Figure 8). Bacterial genomes display a varying degree of duplication. While in the most duplicated genome of *Pseudomonas aeruginosa* >50% of proteins have at least one paralog, in the least redundant *M.thermoautotropicum* the figure is ~10%. The percentage of gene products with at least one significant IMPALA hit to a protein of known structure also varies widely, from 14% in *A.pernix* to 56% in *Mycoplasma genitalium*. *Aeropyrum pernix*, in particular, has the greatest number of structural targets because it is the least structurally characterized and one of the least duplicated genomes.

After the application of STRUDEL the fraction of clustered sequences falls 10-fold, to an average of 3–5% (Figure 9). All sequences still participating in single-linkage clusters are attributed to potential structure determination targets which makes further algorithmic analysis of the clusters unnecessary. The remaining 95–97% gene products are either discarded or end up in the singlet pool and are declared structural targets.

Conclusions

Our procedure automatically yields the minimum set of gene products without any structural homologues and those partially covered by known structural domains. For the latter, structure determination of only individual uncharacterized domains is required. The main observation that we want to demonstrate in this paper is that our pragmatic filtering/re-clustering procedure allows for a dramatic reduction of the number of sequences participating in single-linkage clusters and thus makes the problem of algorithmically rigorous clustering and resolving complex domain similarity problems much less severe. As seen in Figure 1, out of 2235 *E.coli s*equences initially contained in single-linkage clusters, only 127 still remain clustered after the application of the complete target selection procedure.

By default, our algorithm takes as input the complete set of gene products from a given organism. However, the area of application of our technique is not necessarily limited to completely sequenced genomes. The same protocol is suited for any sufficiently large and diverse group of proteins of interest, including proteins known to interact with each other and those involved in a certain cellular process (Terwilliger *et al.*, 1998). STRUDEL has an option to start the analysis with a manually pre-selected protein list.

In this work, we considered as initial targets all predicted soluble proteins possessing substantially large sequence domains without available structural information. All parameters of the analysis are dynamic and can be changed. For example, the choice of the minimum allowed sequence similarity required to join to proteins in a single-linkage cluster depends on the objective of the project. Two proteins sharing a common structural motif at a very low similarity level can be joined in one cluster if the purpose is to obtain a general idea about the folding topology. For detailed studies on structures involving the analysis of individual structural elements, ligand binding sites, etc., a much higher level of homology will be required to join sequences into the same target family.

The decision tree shown in Figure 1 should be considered a rough prototype of the target selection process in a realistic structural genomics project. Each step of the procedure will certainly require further detail. For example, distinguishing between soluble and insoluble proteins is a complex task which goes far beyond mere membrane region prediction. Christendat *et al.* (Christendat *et al.*, 2000) developed a specialized data mining technique for this purpose which involves consideration of hydrophobic stretches in addition to Gln, Asp, Glu and aromatic composition. The process of mapping known three-dimensional structures on genomic sequences should ideally take into account discontinuous domains.

Using the wealth of pre-computed sequence attributes available through the PEDANT database, it is easy to apply a variety of other user-specified criteria for initial screening of gene products that are more likely to yield to expression and crystallization. Those should include protein size and p*I*, the number of cysteine and methionine residues, information on amino acid repeats, predicted exposed surface area and non-globular regions, to name just a few (E.Ulrich, personal communication). Other important features include predicted cellular localization, functional category and the size and phylogenetic distribution of the protein family to which a given protein belongs. Furthermore, the entire body of experimental evidence produced by functional analysis studies (availability of mutants, expression data, protein–protein interactions) should ideally be taken into account. We conclude that the target selection for structural genomics can be best explored in conjunction with extensive high-quality genome annotation.

Acknowledgements

I thank Hans-Werner Mewes for careful reading of the manuscript and useful comments.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) *Nucleic Acids Res.*, **28**, 263–266.

Berman,M.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank Nucl. Acids Res., **28**, 235–242.

Bork,P., Schultz,J. and Ponting,C.P. (1997) *Trends Biochem. Sci.*, **22**, 296–298.

Brenner,S.E., Koehl,P. and Levitt,M. (2000) *Nucleic Acids Res.*, **28**, 254–256.

Bycroft,M., Hubbard,T.J., Proctor,M., Freund,S.M. and Murzin,A.G. (1997) *Cell*, **88**, 235–242.

Christendat,D., Yee,A., Dharamsi,A., Kluger,Y., Savchenko,A., Cort,J.R., Booth,V., Mackereth,C.D., Saridakis,V., Ekiel,I. e*t al.* (2000) *Nat. Struct. Biol.*, **7**, 903–909.

Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E. *et al.* (1998) *Nature*, **393**, 537–544.

Eddy,S.R. (1998) *Bioinformatics*, **14**, 755–763.

Enright,A.J. and Ouzounis,C.A. (2000) *Bioinformatics*, **16,** 451–457.

Fischer,D. and Eisenberg,D. (1999) *Bioinformatics*, **15**, 759–762.

Frishman,D. and Argos,P. (1997) *Proteins*, **27**, 329–335.

Frishman,D. and Mewes,H.W. (1997a) *Nat. Struct. Biol.*, **4**, 626–628.

Frishman,D. and Mewes,H.W. (1997b) *Trends Genet.*, **13**, 415–416.

Frishman,D. and Mewes,H.W. (1999) *Prog. Biophys. Mol. Biol.*, **72**, 1–17.

Frishman,D., Albermann,K., Hani,J., Heumann,K., Metanomski,A., Zollner,A. and Mewes,H.W. (2001) *Bioinformatics*, **17**, 44–57.

Gerstein,M. (1998) *Proteins*, **33**, 518–534.

Hegyi,H. and Gerstein,M. (1999) *J. Mol. Biol.*, **288**, 147–164.

Hwang,K.Y., Chung,J.H., Kim,S.H., Han,Y.S. and Cho,Y. (1999) *Nat. Struct. Biol.*, **6**, 691–696.

Kim,S.H. (2000) *Curr. Opin. Struct. Biol.*, **10**, 380–383.

Kimura,M., Foulaki,K., Subramanian,A.R. and Wittmann-Liebold,B. (1982) *Eur. J. Biochem.*, **123**, 37–53.

Klein,P., Kanehisa,M. and DeLisi,C. (1985) *Biochim. Biophys. Acta*, **815**, 468–476.

Koonin,E.V., Tatusov,R.L. and Rudd,K.E (1996) *Methods Enzymol.*, **266**, 295–322.

Lo Conte,C.L., Ailey,B., Hubbard,T.J., Brenner,S.E., Murzin,A.G. and Chothia,C. (2000) *Nucleic Acids Res.*, **28**, 257–259.

Lupas,A.N., van Dyke,M. and Stock,J. (1991) *Science*, **252**, 1162–1164.

Mallick,P., Goodwill,K.E., Fitz-Gibbon,S., Miller,J.H. and Eisenberg,D. (2000) *Proc. Natl Acad. Sci. USA*, **97**, 2450–2455.

Matsuda,H., Ishihara,T. and Hashimoto,A. (1999) *Theor. Comput. Sci.*, **210**, 305–325.

Milburn,D., Laskowski,R.A. and Thornton,J.M. (1998) *Protein Eng.*, **11**, 855–859.

Moreno-Vivian,C., Cabello,P., Martinez-Luque,M., Blasco,R. and Castillo,F. (1999) *J. Bacteriol.*, **181**, 6573–6584.

Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) *Protein Eng.*, **10**, 1–6.

Park,J. and Teichmann,S.A. (1998) *Bioinformatics*, **14**, 144–150.

Pazos,F., Olmea,O. and Valencia,A. (1997) *Comput. Appl. Biosci.*, **13**, 319–321.

Sali,A. (1998) *Nat. Struct. Biol.*, **5**, 1029–1032.

Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) *Bioinformatics*, **15**, 1000–1011.

Schneider,F., Lowe,J., Huber,R., Schindelin,H., Kisker,C. and Knablein,J. (1996) *J. Mol. Biol.*, **263**, 53–69.

Siomi,H., Matunis,M.J., Michael,W.M. and Dreyfuss,G. (1993) *Nucleic Acids Res.*, **21**, 1193–1198.

Skolnick,J. and Fetrow,J.S. (2000) *Trends Biotechnol.*, **18**, 34–39.

Sonnhammer,E.L.L. and Kahn,D. (1994) *Protein Sci.*, **3**, 482–492.

Storm,C.E. and Sonnhammer,E.L. (2001) *Bioinformatics*, **17**, 343–348.

Terwilliger,T.C., Waldo,G., Peat,T.S., Newman,J.M., Chu,K. and Berendzen,J. (1998) *Protein Sci.*, **7**, 1851–1856.

Wolf,Y.I., Brenner,S.E., Bash,P.A. and Koonin,E.V. (1999) *Genome Res.*, **9**, 17–26.

Wolf,Y.I., Grishin,N.V. and Koonin,E.V. (2000) *J. Mol. Biol.*, **299**, 897–905.

Wootton,J.C. and Federhen,S. (1993) *Comput. Chem.*, **17**, 149–163.

Xu,D. and Nussinov,R. (1998) *Fold. Des.*, **3**, 11–17.

Yokoyama,S., Matsuo,Y., Hirota,H., Kigawa,T., Shirouzu,M., Kuroda,Y., Kurumizaka,H., Kawaguchi,S., Ito,Y., Shibata,T. *et al.* (2000) *Prog. Biophys. Mol. Biol.*, **73**, 363–376.

Yona,G., Linial,N. and Linial,M. (1999) *Proteins*, **37**, 360–378.

Received June 26, 2001; revised November 8, 2001; accepted November 21, 2001