There are several options for viewing an article. You can display Figures in the Text as - Quickest: Links Only, Medium: Thumbnails or Slowest: Full Figures (Current).

**Surprising similarities in structure comparison.**

*Review article*

Jean-Francois Gibrat, Thomas Madej and Stephen H Bryant


**Outline**

- Abstract
- Introduction
- Fast similarity search
- Similarity thresholds
- Refined structural alignments
- Concluding remarks
- Acknowledgements
- References and recommended reading
- Abbreviations
- Copyright

**Abstract**

Examination of a protein’s structural ‘neighbors’ can reveal distant evolutionary relationships that are otherwise undetectable, and perhaps suggest unsuspected functional properties. In the past, such analyses have often required specialized software and computer skills, but new structural comparison methods, developed in the past two years, increasingly offer this opportunity to structural and molecular biologists in general. These methods are based on similarity-search algorithms that are fast enough to have effectively removed the computer-time limitation for structure-structure search and alignment, and have made it possible for several groups to conduct systematic comparisons of all publicly available structures, and offer this information via the
World Wide Web. Furthermore, and perhaps surprisingly given the difficulty of the structure-comparison problem, these groups seem to have converged on quite similar approaches with respect to both fast search algorithms and the identification of statistically significant similarities.

Introduction

Today we know from many examples that three-dimensional (3D) structure is highly conserved in protein evolution [1]. For this reason, comparison of 3D structure makes it possible to peer far back in evolutionary time, and often to make a ‘bridge’ between protein families that appear distinct when judged by sequence comparison alone. The helical cytokines, for example, form an extended family with similar receptors and signaling pathways [2], but their sequences are not detectably similar, and it is only by structure comparison that their common lineage is apparent [3]. Many such extended families are now known, having been discovered by a variety of techniques for structure comparison [4•] [5•] [6•]. Knowledge of these relationships can offer insights as to the properties of an individual family member, by comparison of sequence at functionally important sites, or by use of multiple-structure information in detailed molecular modeling [7]. These similarities also define the conserved structural 'core' of a protein family, information that is critical when attempting to identify distant homologs by fold recognition techniques [8] [9]. For example, a recent [10] and apparently successful [11] prediction that the *Obese* gene product, leptin, is a helical cytokine depended critically on structural alignments spanning this broader protein family.

The interpretation of 3D structure similarity presents a formidable scientific challenge, for it is difficult to distinguish between similarities that arise from evolutionary relationships, with their implications concerning common properties, and similarities that reflect purely physical constraints on protein folding. Extensive sequence similarity implies descent from a common ancestral gene, and the occurrence of many superimposable substructures, preserving the topological connectivity of secondary structure elements (SSEs), provides suggestive evidence, because genetic mechanisms rarely produce topological permutations [12] [13]. Sometimes, however, one observes striking similarities that have nothing to do with common descent. The physical forces driving the formation of β sheets and their packing into 3D layers can lead to extensive 'architectural' similarities [5•] [14], for example, even when the chain ordering of SSEs is completely different. Clearly there is a 'twilight zone', where one cannot determine whether structural similarity arises from biological relationship or physical constraints, or whether substructures resemble one another more than one might expect by chance in any two natural proteins.

Beyond this fundamental scientific question, the detection of 3D structure similarity presents an enormous computational challenge. By whatever metric is chosen, the similarity of 3D structure means the similarity of the relative spatial orientation of many points drawn from each structure, for example, the coordinates of Cα atoms in the polypeptide backbone. But there are a very large number of ways in which one could match backbone atoms from any two proteins, and brute force computation is totally infeasible with today's computers. In the past, researchers have addressed this problem via heuristic algorithms that require either an accurate 'seed' alignment as a starting point [15] [16] [17] [18] [19] or that search for superimposable fragment pairs at the level of backbone coordinates [9] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29]. These programs are, however, either unsuitable or simply too slow for automatic similarity searching and all-against-all 'neighboring' of known structures, particularly given the rapid growth of the Protein Data Bank (PDB) [30]. And yet it is exactly such software tools that are needed if biologists are to routinely explore distant evolutionary relationships. They are also what is needed, in fact, by specialists in the area who wish to conduct further research into the structural 'signatures' of common descent and/or common folding mechanisms.

Appreciating that speed is of the essence in structural comparison, during the past two years, several groups
have endeavored to construct fast search algorithms, to allow rapid 'neighboring' and systematic comparison of
known structures. These methods are based on each group's best understanding of what constitutes interesting similarity, and of the constraints imposed by computer-time
limitations. Surprisingly, given the difficulties mentioned above, these groups have all adopted rather similar
methods with respect to both fast search algorithms and statistical significance thresholds. Borrowing from the
work of Artymuik and colleagues, these methods all employ a high-level representation of protein
structure as an assembly of linear SSEs. Substructure similarities are identified by the types, relative
orientations and chain ordering of SSEs, and residue-by-residue alignments refined in a subsequent step that
need be applied to only the most interesting. Most of these methods also define the statistical significance of a
substructure similarity as a function of the context in which it is observed, and tend to dismiss as 'noise'
substructure similarities that represent only small fragments of a protein or protein domain. In this review, we
summarize this recent work, concentrating on the similarities of these new methods, and presenting examples of
their results. We also point out technical issues still remaining and the sensitivity limitations that one may
foresee.

Of perhaps most interest to the biologist, we also list World Wide Web sites of the structural information
services now offered by several of the groups working in the area of structural comparison and analysis. To see how the new search methods perform, the reader may, in fact, wish to skip ahead to
Table 2, choose a site, and 'click' on a protein of interest. Aside from a better understanding of the new
algorithms, beyond what we can provide in this review, the reader may discover something more important: an
unsuspected similarity to other proteins, and perhaps new insight into the evolution or biological function of an
extended protein family.

### Fast similarity search

The new similarity-search algorithms are based on rapid identification of elementary fragment-pair similarities
consisting of pairs of aligned helices or β strands. A unit of structural similarity thus consists of a pair of SSEs whose type and relative orientation are similar to within a specified
tolerance. This definition contains the germ of protein tertiary structure, but at the same time allows elementary
similarities to be identified rapidly. By comparison of pre-computed internal coordinates, usually distances
and/or angles describing the relative orientation of the axial vectors of SSEs, the fragment pairs that
superimpose poorly may be detected and eliminated from further consideration with only a few computer
operations. The fragment pairs that remain superimpose to the specified similarity and, furthermore contain all
of the potential starting points for a more extensive structural alignment, because the search across all
fragment-pair alignments has been comprehensive. This procedure is similar to the identification of
fragment-pair similarities at the level of backbone coordinates in earlier comprehensive search methods. One gains a tremendous increase in speed, however, by simply defining
the units of similarity to be pairs of superimposable SSEs. There are relatively few SSEs, and
because their internal structures are constrained by hydrogen bonding, one need spend no computer time at all in
the evaluation of intrafragment similarity.

The new algorithms are also similar to one another in the way they assemble aligned SSE pairs into
self-consistent groups that represent a larger, superimposable substructure. A substructure containing three
helices, for example, is identified whenever all of the fragment-pair alignments involving three helices from
each protein meet the specified similarity tolerance. To find such substructures one may simply begin with any
of the SSE pairs, and search down the remaining list of elementary similarities for SSE pair alignments that
contain any of the helices already aligned. This search may be continued, to find four-helix substructures, for
example, and then stopped when no further SSEs may be added. Although there are many possibilities to
consider, this search is quite feasible for the numbers of SSEs found in proteins, when one employs proper bookkeeping to avoid duplicate computation. It is even quite fast when one restricts one’s attention to substructure alignments of potential evolutionary homologs, with conserved chain ordering of the SSEs [31••] [32••] [33••], or considers only SSE pairs in close contact [34••] [35••] [36••]. The ‘clique detection’ algorithm proposed for this problem by Artymuik and colleagues [37] [38] accomplishes this search using an algorithm borrowed from graph theory [44], and it is used in several of the new methods [31••] [33••] [36••]. The remaining methods [32••] [34••] [35••] employ related clustering algorithms, also based on the fast traversal of the list of SSE pair alignments. The result, in all cases, is a comprehensive set of substructure alignments that defines the structural similarity of the proteins under comparison, with search times of only a few seconds or fractions of a second per pairwise protein comparison.

One might, at first, imagine that there will be only a few superimposable substructures shared between any two proteins, but, as the groups working in this area have found, proteins often contain internal symmetries and/or repeats of common supersecondary structure elements. The first few strand/helix units of one TIM (triose phosphate isomerase)-barrel protein [45] [46], for example, may superimpose well on the last strand/helix units of another, even though a global superposition of the entire eight-strand/helix barrel will not contain this alignment. In comparing two large proteins, this similarity of supersecondary structure motifs can generate thousands of alternative, competing structural alignments, even when chain ordering of SSEs is conserved, and one cannot a priori say which is most interesting. To sort through such alternatives, the new algorithms all adopt what is perhaps their most important innovation, scoring schemes that identify the 'best' substructure alignment very rapidly. In some algorithms 'best' is simply defined as the largest substructure that meets the similarity tolerance employed in fragment-pair identification [31••] [32••] [35••]. Other methods take account of both the extent and degree of similarity, measuring the internal coordinate differences [34••], root mean square (rms) Cartesian superposition residuals of SSE axes [33••], or other properties [36••] of the underlying fragment-pair alignments. None of the scoring schemes requires lengthy calculations, however, such as the optimization of residue-by-residue alignments (see below) or the Cartesian-space superposition of all the alternatives. Instead, they identify the 'best' substructure alignment with little additional computation, as fast as alternatives are generated by the search algorithm.

Although the features emphasized above (rapid identification of SSE pair alignments, clustering into groups, and scoring of the 'best' substructure alignment) are common to the new search algorithms, these methods still differ in many details which may yet prove important. The number and boundaries of SSEs within the database structures are defined differently, for example, as are the thresholds of internal coordinate similarity that trigger identification of superimimposable fragment pairs. These definitions may be quite critical when it comes to 'twilight zone' similarities, and certain structural similarities may be missed altogether if SSE pairs happen to exceed the particular tolerance employed. Similarities of small proteins with few or no SSEs may not be detected at all [32••] [34••]. The scoring schemes used to define the best substructure alignment also differ, and certainly may lead to alternative structural alignments. From the examples presented [31••] [32••] [33••] [34••] [35••] [36••], it seems, on the whole, that these new search algorithms perform similarly with respect to the identification of potential homolog structures, where chain ordering among aligned SSEs is conserved. These methods and the 'neighbor' lists they produce are all rather new, however, and a complete analysis remains to be undertaken.

### Similarity thresholds

Having identified the 'best' substructure alignment between two proteins, one is faced with a more difficult question: is this similarity surprising in any way, and perhaps indicative of a biological relationship, or is it simply what one might expect to see by chance? In comparisons of randomly chosen structures it is certainly
rare to see global similarities, where nearly all their SSEs are superimposable and occur in the same chain order. Indeed, such similarity is often accompanied by other evidence of evolutionary relationship, such as sequence similarity or common locations of active sites. Superposition of a few strands or helices, on the other hand, is quite common in random comparisons, and may indicate nothing more than the fact that both proteins contain many such elements, some of which, in compact structures, must necessarily fall in similar orientations [47] [48]. The chance occurrence of similar substructures seems even more likely when one considers that natural proteins fall into only a few architectural classes, each with characteristic supersecondary structure motifs [49]. The purpose of a similarity threshold is to somehow draw a line between these two regimes, to find a similarity score cut-off that determines whether the best substructure alignment identified by a search algorithm is unusual and/or surprising, and to do so in a way that is compatible with fast, automated searching. This is a difficult problem, which may ultimately have many solutions, depending on the nature of the structural similarity one seeks to detect [5•] [32••] [34••].

As surprising as their choice of a common search algorithm is, the groups addressing the structural similarity problem seem to have also converged on a common strategy with respect to definition of significance thresholds. The outline of their approach is to consider the protein context in which a superimposable substructure is observed. If this constitutes a near-global structural alignment, involving most SSEs in either protein, then it will, in general, be deemed significant. The fundamental statistical idea here is that of the size of the search space, the number of alternative substructure alignments that are possible between the two proteins from which one has chosen the best. In a comparison of any two proteins, there are relatively few ways in which one could draw a substructure alignment that equivalences nearly all strands and helices, and it is surprising if any one of them achieves a very good superposition score. In a comparison of two large proteins, however, there are very many ways in which one could draw substructure alignments involving, say, half or a third of their SSEs. This number may easily reach into the millions in practice, and it becomes much less surprising if one substructure alignment achieves a 'one in a million' superposition score, which is just what one would expect to happen by chance. This statistical concept is the same as that that is used in modern methods of sequence comparison. A common 'word' of a few amino-acid symbols is quite surprising in comparison of two short sequences, but the same similarity is indistinguishable from false positives in a database search, where the same word will occur by chance in many sequences [50].

### Table 1. Results of a VAST search for structures similar to a mutant form of bacteriophage T4 glutaredoxin.

<table>
<thead>
<tr>
<th>Fig ID</th>
<th>PDB ID</th>
<th>Chn</th>
<th>Dom</th>
<th>-log(p)</th>
<th>score*</th>
<th>SSE align†</th>
<th>SSE prt-1†</th>
<th>SSE prt-2§§</th>
<th>RES align##</th>
<th>RMSD***</th>
<th>Description of proteins to 1ABA, Glutaredoxin With Val15 Replaced</th>
</tr>
</thead>
<tbody>
<tr>
<td>1AAZ</td>
<td>A</td>
<td>12.9</td>
<td>17.3</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>87</td>
<td>0.9</td>
<td>Glutaredoxin</td>
</tr>
<tr>
<td>1EGO</td>
<td></td>
<td>9.2</td>
<td>12.9</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>62</td>
<td>2.2</td>
<td>1.5</td>
<td>Glutaredoxin (Oxidized) (Structures)</td>
</tr>
<tr>
<td>1GNE</td>
<td>1</td>
<td>8.1</td>
<td>11.6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>68</td>
<td>2.2</td>
<td>1.5</td>
<td>Glutation S-Transfe (E.C.2.5.1.18) Fus</td>
</tr>
<tr>
<td>1GSE</td>
<td>A</td>
<td>8.0</td>
<td>11.5</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>60</td>
<td>2.1</td>
<td>1.5</td>
<td>Mol id: 1;Molecule Glu Transferase</td>
</tr>
<tr>
<td>1GRX</td>
<td></td>
<td>7.9</td>
<td>11.4</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>59</td>
<td>2.3</td>
<td>1.5</td>
<td>Glutaredoxin Mutant Wit Replaced By Ser</td>
</tr>
<tr>
<td>2GST</td>
<td>A</td>
<td>7.6</td>
<td>11.2</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>53</td>
<td>1.8</td>
<td>1.5</td>
<td>1.5</td>
<td>Glutation S-Transfe (Isoenzyme 3-3)</td>
</tr>
<tr>
<td>1GTA</td>
<td>1</td>
<td>7.6</td>
<td>11.1</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>67</td>
<td>2.1</td>
<td>1.5</td>
<td>1.5</td>
<td>Glutation S-Transfe (E.C.2.5.1.18) (26 K)</td>
</tr>
<tr>
<td>1GNE</td>
<td>1</td>
<td>7.5</td>
<td>11.6</td>
<td>7</td>
<td>7</td>
<td>12</td>
<td>66</td>
<td>2.1</td>
<td>1.5</td>
<td>1.5</td>
<td>Glutation S-Transfe (E.C.2.5.1.18) Fus</td>
</tr>
<tr>
<td>Entry</td>
<td>pI</td>
<td>pH</td>
<td>MW</td>
<td>Mol wt</td>
<td>類型</td>
<td>说明</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>--------</td>
<td>------</td>
<td>------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1GSS</td>
<td>7.4</td>
<td>10.7</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>51</td>
<td>1.5</td>
<td>Glutatione S-Transferase (E.C.2.5.1.18)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HNA</td>
<td>7.3</td>
<td>10.4</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>52</td>
<td>1.4</td>
<td>Glutatione S-Transferase Class Mu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1GSE</td>
<td>7.3</td>
<td>11.5</td>
<td>7</td>
<td>7</td>
<td>13</td>
<td>59</td>
<td>2.0</td>
<td>Mol id: 1; Molecule: Glutatione S-Transferase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1GUH</td>
<td>7.2</td>
<td>10.9</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>59</td>
<td>2.1</td>
<td>Glutatione S-Transferase (E.C.2.5.1.18)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1AGS</td>
<td>7.2</td>
<td>11.0</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>58</td>
<td>1.9</td>
<td>Alpha Glutatione S-Transferase (E.C.2.5.1.18)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1GSQ</td>
<td>7.1</td>
<td>10.4</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>53</td>
<td>1.5</td>
<td>Glutatione S-Transferase (E.C.2.5.1.18)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2GST</td>
<td>7.1</td>
<td>11.2</td>
<td>6</td>
<td>7</td>
<td>12</td>
<td>53</td>
<td>1.8</td>
<td>Glutatione S-Transferase (Isoenzyme 3-3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1GTA</td>
<td>7.0</td>
<td>11.1</td>
<td>7</td>
<td>7</td>
<td>12</td>
<td>68</td>
<td>2.2</td>
<td>Glutatione S-Transferase (E.C.2.5.18) (26 Kd)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1GSS</td>
<td>6.8</td>
<td>10.7</td>
<td>6</td>
<td>7</td>
<td>12</td>
<td>51</td>
<td>1.5</td>
<td>Glutatione S-Transferase Class Mu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1AGS</td>
<td>6.8</td>
<td>11.0</td>
<td>7</td>
<td>7</td>
<td>14</td>
<td>58</td>
<td>1.9</td>
<td>Glutatione S-Transferase (E.C.2.5.1.18)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HNA</td>
<td>6.8</td>
<td>10.4</td>
<td>6</td>
<td>7</td>
<td>10</td>
<td>52</td>
<td>1.4</td>
<td>Glutatione S-Transferase Class Mu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1GLQ</td>
<td>6.7</td>
<td>10.2</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>51</td>
<td>1.5</td>
<td>Glutatione S-Transferase (Class Pi)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1GUH</td>
<td>6.7</td>
<td>10.9</td>
<td>7</td>
<td>7</td>
<td>13</td>
<td>60</td>
<td>2.1</td>
<td>Glutatione S-Transferase (E.C.2.5.1.18)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1GSQ</td>
<td>6.7</td>
<td>10.4</td>
<td>6</td>
<td>7</td>
<td>11</td>
<td>53</td>
<td>1.5</td>
<td>Glutatione S-Transferase (E.C.2.5.1.18)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1THO</td>
<td>6.5</td>
<td>10.6</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>57</td>
<td>2.1</td>
<td>Glutatione S-Transferase (E.C.2.5.1.18)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1GLQ</td>
<td>6.1</td>
<td>10.2</td>
<td>6</td>
<td>7</td>
<td>13</td>
<td>50</td>
<td>1.5</td>
<td>Glutatione S-Transferase (Class Pi)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1GSR</td>
<td>6.1</td>
<td>10.1</td>
<td>6</td>
<td>7</td>
<td>13</td>
<td>52</td>
<td>1.5</td>
<td>Glutatione S-Transferase (E.C.2.5.1.18)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2TIR</td>
<td>6.1</td>
<td>10.1</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>57</td>
<td>2.1</td>
<td>Glutatione S-Transferase (Class Pi)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1SRX</td>
<td>6.0</td>
<td>9.2</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>47</td>
<td>1.9</td>
<td>Thioredoxin (Oxidized)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1TRX</td>
<td>6.0</td>
<td>10.5</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>58</td>
<td>2.4</td>
<td>Thioredoxin (Reduced)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3TRX</td>
<td>5.7</td>
<td>9.4</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>39</td>
<td>1.9</td>
<td>Thioredoxin Mutant With Replaced By Ala</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1TRS</td>
<td>5.7</td>
<td>9.4</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>42</td>
<td>2.0</td>
<td>Thioredoxin Mutant With Replaced By Ala</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1MDI</td>
<td>5.2</td>
<td>8.9</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>43</td>
<td>2.2</td>
<td>Dsba (Disulfide Bond For Protein)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1DSB</td>
<td>5.1</td>
<td>8.8</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>37</td>
<td>1.8</td>
<td>Dsba (Disulfide Bond For Protein)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>Seq. No.</td>
<td>Mol. ID</td>
<td>Mol. Name</td>
<td>VAST p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>----------</td>
<td>-----------</td>
<td>------------------------------------</td>
<td>--------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1DSB</td>
<td>4.6</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1DIH</td>
<td>4.5</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2ATC</td>
<td>4.4</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1GP1</td>
<td>*3.9</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1COY</td>
<td>2.9</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1PBF</td>
<td>2.8</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1PBD</td>
<td>2.8</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1PXC</td>
<td>2.3</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1PXB</td>
<td>2.3</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1DBP</td>
<td>3.4</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1DBP</td>
<td>1.4</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1PXA</td>
<td>2.8</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1PVD</td>
<td>2.5</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1PYD</td>
<td>2.5</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1PHH</td>
<td>2.2</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2GLT</td>
<td>3.2</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1GLV</td>
<td>3.2</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1DRK</td>
<td>3.0</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2DRI</td>
<td>3.0</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1DRJ</td>
<td>3.0</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2LIV</td>
<td>2.7</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2LIV</td>
<td>2.7</td>
<td>50</td>
<td>2.9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The figure in this review that shows the superposition. † Protein Data Bank structure code. ‡ Protein Data Bank code. § Arbitrary compact domain identifier. # VAST p-value. An asterisk in this column indicates -log(p)<4, the significance threshold. Hits above this value are sorted by -log(p), hits below by SSE superposition score. ** VAST superposition score. †† Number of SSEs aligned by VAST. && Number of SSEs in the query structure. §§ Number in the similar structure. ## Number of residues in refined VAST alignment. *** Cα superposition residual of refi
Though they employ the same statistical concept, the new structural similarity search methods nonetheless differ in their detailed definition of significance thresholds. Some methods adopt an empirical approach, comparing the superposition score obtained in a particular match to the distribution of scores for randomly chosen proteins of a similar size [32••] [35••], because the search space for alternative SSE alignments is largely a function of protein size. Other methods simply adjust their superposition score according to the fraction of residues that are aligned, so that structural alignments involving only a portion of either structure will necessarily get poor scores [26] [29] [31••]. The VAST method (vector alignment search tool) [33••] takes account of search space size in a way that is perhaps most analogous to sequence comparison methods, such as its namesake, the BLAST (basic local alignment search tool) algorithm [51]. VAST calculates a p-value for the best substructure superposition as the probability that this score would be seen by chance in drawing SSE pairs at random, multiplied by the number of alternative substructure alignments possible given the SSEs in the protein pair under comparison. The latter value is the size of the search space, which may be calculated explicitly when structures are represented as a set of distinct, non-overlapping SSE fragments (J-F Gibrat, T Madej, JL Spouge, SH Bryant, unpublished data). The p-value calculation makes use of an empirical distribution of superposition scores for randomly aligned fragment pairs, in place of BLAST's analytical result for random substring alignments, and the search space is determined by a combinatorial formula giving the number of possible SSE alignments, rather than a simple sequence-length difference, but the statistical theory is identical. A substructure superposition achieving a score expected in only one in $10^8$ random substructure alignments, for example, as the best among $10^4$ possible substructure choices, is significant at $p = 10^{-8} \times 10^4 = 10^{-4}$.

The results of context-dependent significance thresholds are perhaps best judged by example. Table 1 shows the results of a VAST search for structures similar to a mutant form of bacteriophage T4 glutaredoxin, PDB code 1ABA [52]. The table lists p-values, substructure superposition scores and other information for all similar structures detected, and for purposes of illustration includes some 'hits' below VAST's normal significance threshold. Substructure superposition scores may be seen to be related to the number of SSEs aligned, the number of residues aligned by refinement (see below), and their rms residual, as one might expect. Statistical significance of the various hits, however, is seen to also depend on the proportion of SSEs included in the aligned substructure. Those corresponding to a near-global structural alignment, such as the similarity to human thioredoxin, PDB code 1TRS [53], show lower p-values than hits corresponding to a partial substructure alignment, such as Pseudomonas fluorescens p-hydroxybenzoate hydroxylase, PDB code 1PBF [54]. The latter similarity is judged to be insignificant by the context-dependent threshold, even though it aligns nearly as many residues to a better rms residual. The SSE alignments for these two examples are shown in Fig. 1, where the larger search space in the 1ABA-1PBF comparison is apparent from the larger number of SSEs in 1PBF. In searching a database assumed to contain 500 independent structures, a p-value of $10^{-4}$ corresponds to a conventional significance threshold of $p \leq 0.05$ with an appropriate correction for database size, $p \leq 0.05/500 = 10^{-4}$ [33••] [50]. In this example, most hits above this threshold are glutaredoxins, thioredoxins, and glutathione S-transferases, proteins which share a common function in the catalysis of disulfide oxidation/reduction, even though only one, a non-mutant form of T4 glutaredoxin, shows significant sequence similarity by BLAST [51]. Context-dependent significance thresholds take into account only a statistical effect, but as this example shows, the similarities most surprising from a statistical point of view are often those which are most interesting biologically.
Figure 1 Examples of structural alignments. (a) Structural alignments for a similarity above the VAST context-dependent significance threshold, T4 glutaredoxin (PDB code 1ABA) versus human thioredoxin (PDB code 1TRS), and (b) a similarity below this threshold, T4 glutaredoxin (PDB code 1ABA) versus P. fluorescens p-hydroxybenzoate hydroxylase (PDB code 1PBF). T4 glutaredoxin is shown in green, as a Cα backbone trace, and with axial vectors of its SSEs shown as half cylinders (helices) and bars (strands). A ligand binding in its active side, N-ethylsulfite morpholine, is shown in stick form in the center (gray, blue, red and yellow). SSEs for the structural neighbors of glutaredoxin are shown in yellow. VAST’s SSE alignment in both cases includes the four β strands 'below' the active-site ligand, and the helix appearing to its right. The SSE alignment of glutaredoxin and thioredoxin also includes the helix appearing to the left of the ligand. Figure generated using the CN3D viewing software used by Entrez [65].

Return to Figure reference in text 1, 2, 3, 4, 5, 6.
Figure 2 Refined structural alignments for T4 glutaredoxin (PDB code 1ABA) versus E. coli disulfide bond formation protein (PDB code 1DSB-A), as determined by (a) VAST, and (b) DALI. T4 glutaredoxin is shown in green, with its active-site ligand in stick form in the center (gray, blue, red and yellow). The disulfide bond formation protein is shown in yellow. Cα backbone traces are shown for both proteins, with residues included in the refined alignments highlighted by thickening. Alignments are taken from the Entrez [42] (a) and FSSP [41] [59] (b) databases. They superimpose 50 residues to an rms residual of 2.9 Å in the case of VAST, and 72 residues to an rms residual of 3.9 Å in the case of DALI. Figure generated using the CN3D viewing software used by Entrez [65].

Return to Figure reference in text 1, 2.
The use of a context-dependent similarity threshold has important implications with respect to the 'modules' of protein structure that one should consider in a database search. The structures of many immune-system proteins, for example, contain several repeats of the immunoglobulin 'fold', an antiparallel β sandwich of roughly 110 residues. In comparing such large structures with the structure of a single-domain fragment, a context-dependent threshold may well indicate that the similarity is insignificant, because it extends over only a small fraction of the larger protein. To achieve the best possible sensitivity, most groups thus conduct comparisons at the level of protein modules, or domains, either by explicitly parsing structures into compact units via heuristic algorithms, or by using superposition scores weighted only by the size of the smaller of two proteins compared. Many of the VAST hits in Table 1, for example, involve compact domains identified within glutathione S-transferase structures, and they are more significant than the corresponding hits involving the complete structure. While they account for the modular nature of proteins in these ways, and base their comparison on the similarity of modules, it seems likely that all of these methods can be improved by the incorporation of more precise definitions of domains, using the results of prior sequence or structural analysis.

One should also bear in mind that the similarity thresholds employed by these methods allow only for the increased chance of finding a similar substructure when one has more possibilities to choose from; they do not determine whether an observed similarity arises from evolutionary relationship or convergence to a favorable folding motif. The occurrence of a similar strand-helix-strand substructure in two proteins may be judged insignificant, but two occurrences of this substructure may cross the threshold and be judged significant. Does this mean that the two proteins are homologs, or only that this architectural feature helps proteins to fold rapidly, and is a likely point of convergence? The questions of statistical significance and biological relationship are not unrelated, because a structural similarity with a high probability of occurring by chance, in the context of a particular comparison, provides no evidence for a biological relationship. But the converse is not true. Certain 'superfolds' show highly significant structural similarity, although the question of their evolutionary origins remains open.

Refined structural alignments

The new fast search algorithms produce an alignment of secondary structure elements, as was illustrated in Fig. 1. A corresponding residue-by-residue alignment is implied, since the atomic structures of strands and helices are similar to one another, but further computation is still required to find a favorable alignment of backbone-atom coordinates. SSE alignments must therefore be subjected to a refinement procedure that considers various alternative residue-by-residue alignments, ranks them, and chooses the best, or perhaps the suboptimals above a specified tolerance. Fortunately, this computational problem is nearly identical to that of refining an initial 'seed' alignment, the question first addressed many years ago in pioneering work in structural alignment, and improved on subsequently. The new algorithms can use updated versions of these refinement algorithms, with initial residue seed taken to be any of those consistent with overlap of the aligned SSEs.

The refinement algorithms used by the new methods fall into two broad classes, iterative superposition and Monte Carlo (MC) optimization, due to differences in interatomic distances. In iterative superposition methods, the seed alignment is used to perform a Cartesian-space superposition. A dynamic programming method is then used to identify the best-aligning residue pairs, which form the basis of the next superposition. Because dynamic programming algorithms require independent scoring of aligned residue pairs, these methods generally employ a scoring function that accepts all aligned-residue pairs falling under a fixed superposition threshold, such as 3 or 4 Å. These algorithms are very fast, but may retain any poorly aligned fragments present in the initial seed. MC algorithms iteratively explore a series of shifts in the alignment...
of individual fragments, and extension by addition of new aligned residue pairs [21] [27] [33••] [61]. The alignment scores employed increase with the number of residues aligned, and decrease as rms residuals worsen, seeking to balance the number of residues aligned with the increase in rms one can expect for longer alignments, but they differ in detail [32••] [33••]. MC refinement methods require more computer time, with refinement of VAST alignments for significant hits, for example, typically taking as long as the fast search for those hits. We note that search methods based on fragment-pair superpositions at the level of backbone coordinates directly refine residue-by-residue alignments [9] [20] [22] [23] [24] [25] [26] [27] [28] [29], and it seems likely that any of these methodologies, with their well-researched scoring schemes, could be adapted to the refinement of the 'seed' alignment determined by a fast search method.

It is difficult to judge whether differences in the methods of refining a detailed structural alignment are of great importance. Results can differ slightly, as shown in Fig. 2 by the alternative alignments of T4 glutaredoxin, PDB code 1ABA [52] with *Escherichia coli* disulfide bond formation protein, PDB code 1DSB-A [62]. Fig. 1a is taken from the Entrez [42] database, which uses the VAST [33••] algorithm, and Fig. 1b is taken from the FSSP (families of structurally similar proteins) [41] database, which uses the DALI (distance matrix alignment program) [27] [32••] algorithm. Both of these methods use MC refinement, but scoring schemes are based in the first case on global similarity of interresidue distances, and in the second on weighted distance similarities, where the agreement of short-range interactions is emphasized. In this example, VAST produces a lower rms alignment, encompassing fewer residues, and DALI a more extensive alignment, at higher rms. It is unclear that either is 'better' than the other. The VAST alignment might be useful for identification of highly conserved core elements, for example, for a threading experiment. The DALI alignment might be useful if one wishes to identify a larger set of analogous sites, for example, in homology modeling. It is perhaps a good thing that some variety exists at the level of scoring schemes for refinement, because the different structural alignments they produce may, in the end, be suited to different purposes [34••] [63].

---

**Concluding remarks**

These new algorithms provide new information, not so much because they detect similarities that could not in principle be found before, but because they are fast enough to make automated similarity searches involving all structures technically possible. Furthermore, many groups conducting this work have generously made their results available in an easily accessible form, via sophisticated browsing software available over the Internet and World Wide Web. With this development, structure-similarity searching, and the biological insights it can provide, has suddenly become a tool available to most scientists. As has been the case for sequence comparisons, there are discoveries waiting to be made, and they will be made less and less by specialists in structure comparison, and more and more by biologists exploring the sequence and structure 'neighbors' of particular proteins, whose function and evolutionary lineage they wish to understand [64].
Table 2. Web sites for structural similarity databases.

<table>
<thead>
<tr>
<th>Resource</th>
<th>World Wide Web address</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSSP</td>
<td><a href="http://swift.embl-heidelberg.de/">http://swift.embl-heidelberg.de/</a> fssp/</td>
<td>Structure neighbors, structural alignments, similarity search service.</td>
</tr>
<tr>
<td>SCOP</td>
<td><a href="http://scop.mrc-lmb.cam.ac.uk/">http://scop.mrc-lmb.cam.ac.uk/</a> scop/</td>
<td>Structure neighbors, fold classification, 3D viewing.</td>
</tr>
</tbody>
</table>

Return to table reference 1, 2, 3.

In Table 2 we list the World Wide Web addresses of sites that provide access to the results of structural similarity search and classification. These include the results of the new algorithms described here [33••] [35••], and results of other comprehensive similarity search and human expert classifications [39] [40•] [59]. A brief description of the information available at each site is provided in Table 2. These are wonderful resources, well worth a stop on a biologist's tour of the Internet!

Acknowledgements

We thank Stephen Altschul, Chris Hogue, Ina Koch, Hitomi Ohkawa, John Spouge and John Wilbur for valuable discussion and comments on the manuscript. This work was supported by the Intramural Research Program of the National Institutes of Health.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest.
- of outstanding interest.


10. Madej T, Boguski MS, Bryant SH: Threading analysis suggests that the obese gene product may be a helical cytokine. *FEBS Lett* 1995, **373**: 13-18. [MEDLINE], Return to reference citation 1


   Return to reference citation 1, 2

   Return to reference citation 1, 2

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1, 2, 3

   [MEDLINE], Return to reference citation 1, 2

   [MEDLINE], Return to reference citation 1, 2, 3

   [MEDLINE], Return to reference citation 1, 2, 3

   [MEDLINE], Return to reference citation 1, 2, 3

   [MEDLINE], Return to reference citation 1, 2, 3

   [MEDLINE], Return to reference citation 1, 2, 3, 4, 5

   [MEDLINE], Return to reference citation 1, 2, 3, 4, 5

[MEDLINE], Return to reference citation 1, 2, 3


Return to reference citation 1, 2, 3, 4, 5


Return to reference citation 1


This paper outlines a fast search algorithm based on SSE alignment, followed by refinement at the level of backbone coordinates.

[MEDLINE], Return to reference citation 1, 2, 3, 4, 5, 6, 7, 8, 9


A detailed description of the 3D Lookup method, a fast search based on SSE alignment followed by dynamic programming refinement.

[MEDLINE], Return to reference citation 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13


This paper outlines the VAST method, in the context of identifying 'true positives' in a fold-recognition experiment. This employs fast search by SSE alignment, followed MC refinement at the level of backbone coordinates.

[MEDLINE], Return to reference citation 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13


A thoughtful discussion of scoring schemes for SSE fragment pairs, and of potential pitfalls in SSE-based similarity search methods.

Return to reference citation 1, 2, 3, 4, 5, 6, 7, 8, 9


A detailed description of the SARF2 method, based on fast search by SSE alignment, followed by dynamic programming refinement.

Return to reference citation 1, 2, 3, 4, 5, 6, 7, 8, 9, 10


The authors describe a modification of the clique detection algorithm, and use of SSE representations for multiple-structure alignment.

Return to reference citation 1, 2, 3, 4, 5, 6

37. Mitchell EM, Artymuik PJ, Rice DW, Willett P: Use of techniques derived from graph theory to


   [MEDLINE], Return to reference citation 1, 2

   [MEDLINE], Return to reference citation 1, 2

   [MEDLINE], Return to reference citation 1, 2

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1, 2

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1, 2, 3

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1


Abbreviations

BLAST—basic local alignment search tool;
3D—three-dimensional;
DALI—distance matrix alignment program;
FSSP—families of structurally similar proteins;
MC—Monte Carlo;
PDB—Protein Data Bank;
rms—root mean square;
SSE—secondary structure element;
VAST—vector alignment search tool.

Author Contacts

Jean-Francois Gibrat, Unité de Bio-Informatique, Bâtiment de Biotechnologies, Institut National de la Recherche Agronomique, Jouy-en-Josas, 78350, France. Thomas Madej and Stephen H Bryant, National Center for Biotechnology Information, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA.

E-mail address for Stephen H Bryant (corresponding author): bryant@ncbi.nlm.nih.gov

Publication Details

**Publisher:** Current Biology Ltd
London

**Address:** Middlesex House 34-42 Cleveland Street
London