Capri/MR: Exploring Protein Databases from a Structural and Physicochemical Point of View

Eric Paquet
National Research Council Canada
Building Montreal Road
Ottawa, Ontario, Canada
+1 613-991-5035
eric.paquet@nrc-cnrc.gc.ca

Herna L Viktor
University of Ottawa
SITE Building, 800 King Edward Road
Ottawa, Ontario, Canada
+1 613-562-5800 2341
hlviktor@site.uottawa.ca

ABSTRACT

With the advent of high throughput systems to experimentally determine the three-dimensional (3-D) structure of proteins, molecular biologists are in urgent need of systems to automatically store, maintain and explore the vast structural databases that are thus being created. We have designed and implemented the Capri/MR system which makes it possible to identify families of protein structures, as contained in such very large 3-D protein structure databases. Our system is able to automatically index and search a database of proteins by three-dimensional shape, structural and/or physicochemical properties. For each of these diverse protein structure representations, we create a compact rotation and translation invariant index (or signature) which is placed in a database for future querying. A similarity search algorithm performs an exhaustive search against the entire database. Our search algorithm takes advantage of the compact signatures to rapidly find protein structures that are similar in 3-D shape and/or two-dimensional (2-D) properties. As a result, queries in our Capri/MR system run within a fraction of a second, and we are able to accurately group protein structures into the correct families, with very high precision and recall. In addition, our system dynamically processes new protein structures as they become available. We demonstrate the power of Capri/MR against the Protein Data Bank, which contains all known, experimentally determined, 3-D protein structures (48.000 as of January 2008). The main applications of our Capri/MR system lie in structural proteomics, protein evolution and mutation, as well as in drug design, in particular for studying the docking problem and the computer aided design of non-toxic drugs.

PVLDB '08, August 23-28, 2008, Auckland, New Zealand
Copyright 2008 VLDB Endowment, ACM 978-1-60558-306-8/08/08

1. INTRODUCTION

The number of experimentally determined known 3-D protein structures is expected to grow linearly, with an estimated 100 new structures being created every week [1]. This number is indicated to grow exponentially with the advent of high throughput systems. There is an urgent need for systems which will enable molecular biologists to effectively store, manage and explore these vast repositories. That is, domain experts need systems which automatically update their databases as new structures become available. They require systems to determine whether a protein structure is, indeed, new and to which family it appears to belong. Furthermore, they are in want of systems to enable them to accurately find similar structures fast; and to further explore the properties of these similarities, in order to aid them to explain mutations, find similar proteins that appear to have related functionalities, find docking sites, and so on.

We have designed and implemented the Capri/MR system which is able to index and search a very large database of proteins by its three-dimensional shape, structural and/or physicochemical properties. By representing protein structures from different viewpoints, scientists are able to obtain new insights into the structural properties (e.g. for drug design), atomic composition, or local shape (which is useful for e.g. studying the docking problem). They are also able to consider similar amino acid sequences or secondary structures. Our automatic indexing algorithms created compact signatures which are rotation and translation invariant. Our similarity search technique employs a Query by Example (QBE) paradigm, in order to accurately identify those structures that are similar to a query (or so-called seed), when compared to all structures in the database.

In this presentation, we describe the architecture of the Capri/MR system and demonstrate its applicability to the Protein Data Bank. In our Capri/MR system, an exhaustive search is performed in less than a second. Furthermore, the database and the results may be exported and the proteins may be seamlessly visualised interactively in a viewer or in a virtual reality (VR) environment.

This proposal is organized as follows. In Section 2, we describe the architecture of Capri/MR. Section 3 contains the details of the demonstration and Section 4 concludes this demonstration proposal.

2. Capri/MR SYSTEM

The Capri/MR system consists of three main components, as depicted in Figure 1. The first is an algorithm to create different representations of a protein structure. The second component constitutes the signature creation (or so-called indexing) algorithms, which construct 3-D and 2-D signatures of each protein structure. The third component is the similarity search engine that locates protein structures with similar indexes (or signatures), using a Query by Example (QBE) paradigm.

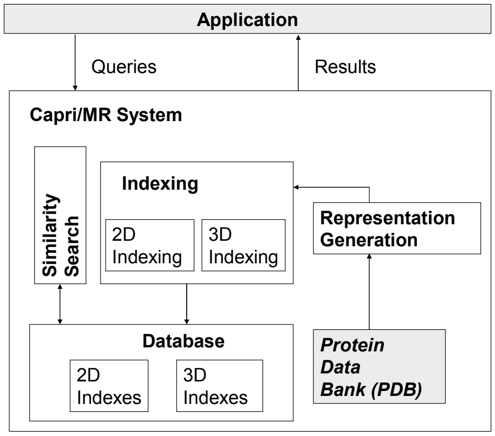


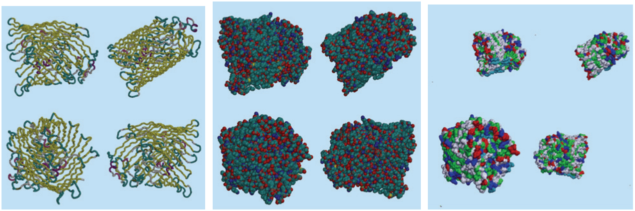
Figure 1: Capri/MR Architecture

2.1 Representation Generation

The representation generation component of our system generates six (6) different representations, as summarized in Table 1. The representations correspond to three-dimensional shape, secondary structures, atomic structure, external three-dimensional appearance (or envelope), amino acid (or residue) organisation and amino acid type organisation.

Table 1: Representations and their properties

Property	Representation	Viewpoint
3-D Shape	Tube Envelope Van der Waal	Backbone, Drug Docking, Drug Atomic Composition
Structural	Secondary Structures (2-D)	Classification, Evolution, Mutation
Physico-chemical	Residue Name (2-D) Residue Type (2-D)	Composition, Evolution, Mutation, Interaction, Drug



(a) Tube (b) Van der Waal (c) Envelope

Figure 2: Three different 3-D representations of Ferric hydroxamate uptake receptor FhuA from Escherichia coli

For each protein, three distinct 3-D representations are associated: the Tube representation which associates a cylindrical tube to the main chain(s) of C-alpha carbons (the back bone of the protein), the Van der Waal (VDW) representation which associates a sphere to each atom forming the protein and the Envelope representation, which corresponds to the outer surface of the protein. Figure 2 (a) to (c) shows these three different 3-D representations, as generated for the Ferric hydroxamate uptake receptor FhuA from the Escherichia coli protein structure.

The Tube representation is important in order to study the structural properties of proteins which are essential, for instance, for protein classification and drug design. Secondly, the Van der Waal representation corresponds to the atomic distribution associated with a given protein. This is the most fundamental representation, in the sense that it does not involve an extended analysis of the experimental data. One should remember that a three-dimensional structure is obtained either through X-ray crystallography or magnetic nuclear resonance (MNR), from both of which the position of the constituent atoms may be inferred. Thirdly, the Envelope representation is particularly relevant for the docking problem. Its importance is related to the fact that the interaction of two proteins is determined, in part, by the local shape of the interacting region or contact zone. If the two 3-D shapes match, the associated proteins may dock and interact. All the 3-D representations mentioned above are important to study possible evolutions, explain mutations and to e.g. replace a toxic protein, with a suitable functionality, by a non-toxic one.

For each three-dimensional representation, one may encode, in 2-D, a certain number of structural and physicochemical properties [2] by representing them with a colour code on the 3-D structure. For the Tube representation, the colour assigned to each cylindrical element is related to the so-called secondary structures of the protein e.g. alpha helix, beta sheet, etc. These secondary structures are of importance for protein classification and to understand their possible evolution and mutations. From the interaction point of view, a colour may be associated either with the amino acids; (there are twenty of them e.g. the ALA is encodes in blue, the LEU in Pink and the HSD in cyan), or with the amino acid types, which group them into eight (8) categories, e.g. a solvent is coded in yellow, an acid in red, a polar in green and an ion in tan. The amino acids are the building blocks of the proteins. They are attached to the back bone as a side-chain and their sequence determines the folding, i.e. the 3-D shape of the corresponding protein. Proteins with similar amino acid sequences are likely to have a similar shape and interact in a related fashion, which is very important in drug design. For example, consider the scenario where two proteins *A* and *B* have similar structures and functionalities, with *A* being toxic and exhibiting serious side effect, while *B* is non-toxic with no or minor side effects. In this case, it is advisable to replace the toxic combination of *A* by that of *B*. This may be done by searching for proteins for which the amino acid sequence(s) are similar.

The next step involves the creation of the signatures, as discussed next.

2.2 Creating the Signatures

Here, we briefly describe our signature creation algorithms. Interested readers are referred to [3] for a detailed discussion.

The 3-D signature (or index) provides a complete description of the 3D shape of the protein. It is not affected by the position or the orientation of the protein in space. This means that the proteins in the database do not need to have a standard orientation or to be aligned relative to one another; a computationally expensive requirement. For each protein, a triangular mesh is associated with the representation from which a tensor of inertia is computed. Subsequently, a reference frame is associated with the Eigen vectors of the later and an invariant statistical distribution (weighted radial and angular) of the triangles of the mesh is calculated. The binary signature obtained is small, only 128 bytes, and its size is independent on the size of the associated protein. This signature describes either: (a) the 3-D structure of the C-alpha chain (the back bone of the protein), (b) the Van der Waal atomic distribution or (c) the envelope (the 3-D outer appearance of the protein) depending on the selected representation.

For the 2-D signature of a given representation, four Eigen 2-D views (projections) of each protein are captured, for which a random morphological analysis is performed in terms of composition and texture [3]. This is achieved by accumulating the relative proportion of colours within a structural element which is moved randomly on each Eigen view. A 256 bytes signature, independent of the size of the protein, is generated. This signature encodes either (a) the description of the structures, (b) the names of the amino acids or (c) the amino acid types, depending on the representation created.

2.3 Searching for Similar Structures

Our similarity search algorithm employs a Query by Example (QBE) paradigm in order to find similar protein structures. That is, a query protein structure is used as a “seed” and the most similar structures are then located. For example, let us assume that we want to calculate the similarity of all proteins in PB_r, i.e. all proteins presented using representation *r* against a query protein P_{q1}. We calculate the similarity measure between P_{q1} and each other protein structure in PB_r. This distance is given by the Euclidian distance

$$d = \sqrt{\sum_{i=1}^{N} \left(x_i - y_i\right)^2} \qquad (1)$$

where *i* and *j* denote the various dimensions, *x* is the signature associated with the unknown structure and *y* is the target structure. Note that an exhaustive search is performed; it will be demonstrated that our Capri/MR system search component is very fast due to the compact nature of our signatures. Also, the query protein structure does not have to be part of the database, i.e. it may be a new protein structure submitted by a user. In such a case, the representations and signatures are generated automatically without any human intervention; the process is completely transparent to the user. The first *N* results are then displayed as thumbnails, where *N* is user defined. For each thumbnail, four views of the associated protein are shown. Then, it is possible to select a particular protein and to obtain the associated metadata as well as the associated signature which is displayed as a chain of numbers. It is also possible to access the file associated with the protein and to visualise the later in three-dimensions interactively, in order for instance, for an expert to examine a certain feature in detail. The system also allowed transferring automatically the protein file to our virtual theatre [4] in which the protein can be visualised in stereo in a VR environment, in order to enhance the interactivity as well as to facilitate collaborative work. In addition, it is possible to save the results, including the signatures, and to export it in a flat file to subsequently process them using other software, such as a data mining package, in order to perform classification. In addition, the search may be re-iterated from any result, in order to further explore similar structures.

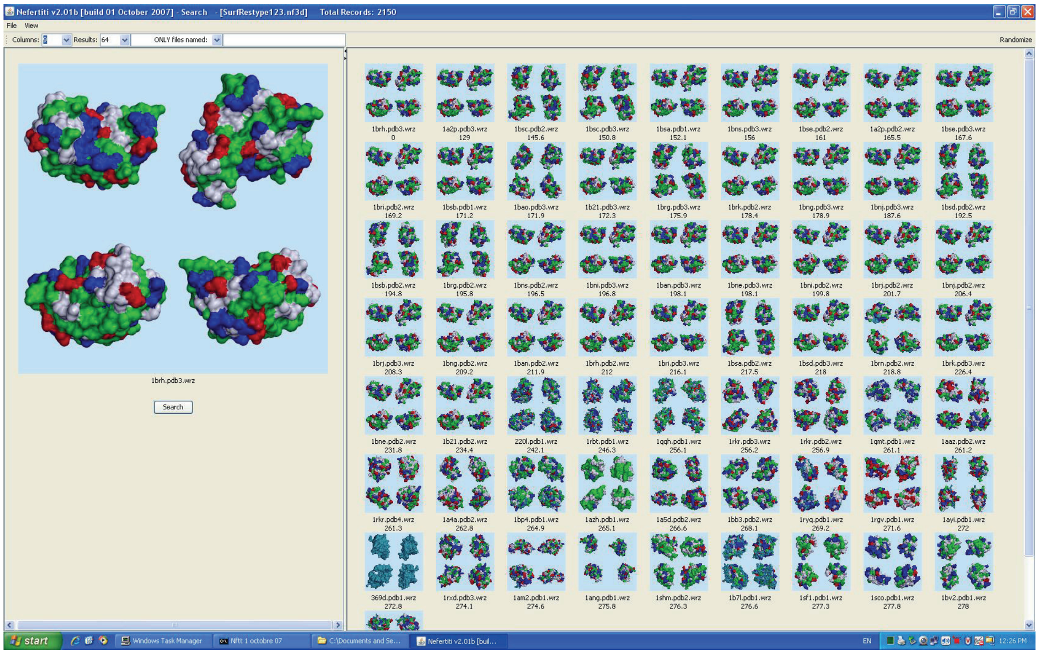
3. DESCRIPTION OF DEMONSTRATION

In this demonstration, we will illustrate the following central features of the Capri/MR system:

1. The ability to create six diverse representations of a protein structure.
2. The automatic, transparent creation of 3-D and 2-D signatures, based on these six representations.
3. The fast, accurate retrieval of protein families of different sizes and appearances; and the location of inter-family similarities. Capri/MR performs an exhaustive search against a very large database, in a fraction of a second.
4. The usefulness of the various representations and their applicability in structural proteomics, in particular in drug design and when studying the docking problem.

3.1 Demonstration Setup and Result

We illustrate the performance of the Capri/MR system against the Protein Data Bank (PDB), which contained 48.000 different 3-D protein structures in January 2008 [5]. All our results are verifiable by using the SCOP (Structural Classification of Proteins) system that describes the structural relationships of proteins of known structure [6]. In the SCOP classification system, proteins are grouped into families, based on experts’ experience. More specifically, proteins are classified into (from large to small) folds, super-families and families. As such, it provides us with the experts’ evaluation of our query results. The analysis of our system’s accuracy and performance indicate that we are able to accurately locate protein families, with a high precision and recall rate [3]. That is, our system is able to locate a query protein structure’s family; and also indicate related families when performing an exhaustive search against 45.000 proteins. For example, Figure 3 shows the results obtained when using the `1brh` structure from the Bacterial Ribonucleases family as query, when utilizing the 3-D envelope representation. The figure shows that Capri/MR system was able to locate the family members (with a precision of 100% (39/39) and a recall of 100%) when using the `1brh` structure as seed. As another example, our system is able to find the family members of the 95 member Homo Sapiens Hemoglobin using the `1rly` as query structures and the 3-D Tube representation, with very high precision and recall. Namely, the first 55 similar structures retrieved belong to the Homo Sapiens Hemoglobin protein structure family; with 86 of the first 100 structures retrieved belong to this family, i.e. a precision of 86% was obtained with a recall of 90%. Interestingly, the presence of Hemoglobin structures from other species, such as a cow (position 56), rookcod (position 69) and a chicken (position 71) convince that we are able to find important inter-family similarities as well [3]. Similar results hold for other families within the PDB when using the four (4) other protein representations, as will be illustrated during our demonstration.



Nefertiti v2.01b [build 01 October 2007] - Search - [SurfRestype123.nf3d] Total Records: 2150

Columns: 9 Results: 64 ONLY files named:

Figure 3: Capri/MR Search Results for the Bacterial Ribonucleases family using the `1brh` structure as query with the 3-D envelope representation

The Capri/MR system has been developed in Java and Java 3D and the calculated signatures are stored in an IBM DB2 V9.1 database. Capri/MR may operate either from Linux or Windows OS and runs from a USB key without any installation (it has its own virtual machine). This implies that it may be operated on-site with minimum privileges, an important security feature for many pharmaceutical companies. For the computation of the signatures, we used a 32 bits Windows XP workstation with two Intel Xeon™ processors, 8 GB of memory and two high-end *n*VIDIA Quadro™ graphical processing units. These are, by no means, minimum requirement, since the querying process may be easily performed on a portable computer, even on a tablet, as will be illustrated during our demonstration.

SUMMARY

In this presentation, we describe the design and implementation of the Capri/MR system. The main applications of our system are in structural proteomics, protein evolution and mutation and drug design, in particular for studying the docking problem and the computer aided design of non-toxic drugs.

Our Capri/MR system is appealing to both experts and the neophytes in the sense that highly technical and complex information can be extracted from the database by performing simple, intuitive and visual queries. This is due to the fact that most of the complexity has been hidden in the signatures, which are searched transparently using a QBE approach. Because of their inherent structure and compact support, the signatures allow us to perform an exhaustive search in the database. This is in contrast to other current approaches [7], which have to rely on a heuristic method in order to perform the query in a reasonable amount of time. In addition, our Capri/MR automatically creates signatures for new structures and incorporates them seamlessly into the existing database. Future work will include a robust comparison to other state-of-the-art techniques as well as a study to determine its usability from a molecular biologist's perspective.

3. ACKNOWLEDGMENTS

This research was partially supported by the Canadian Foundation of Innovation (CFI) and the Canadian National Science and Engineering Research Council (NSERC).

4. REFERENCES

[1] J.-S. Yeh, D.-Y. Chen and M. Ouhyoung, A Web-based Protein Retrieval System by Matching Visual Similarity, *Bioinformatics*, 21(13), pages 3056-3057, 2005.

[2] A. M. Lesk, *Introduction to Protein Science: Architecture, Function, and Genomics*, Oxford University Press, 2004.

[3] E. Paquet and H. L. Viktor, Exploring Protein Architecture using 3D Shape-based Signatures, *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1204-1208, 2007.

[4] E. Paquet and H. L. Viktor, Distributed Virtual Environments for Visualization and Visual Data Mining, *ISPRS Int. Workshop on "Visualization and Animation of Reality-based 3D Models"*, 6 pages, CD ROM, 2003.

[5] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, The Protein Data Bank, *Nucleic Acids Research*, 28, pages 235-242, 2000.

[6] A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, SCOP: A Structural Classification of Proteins Database of the Investigation of Sequences and Structures, J*ournal of Molecular Biology*, 247, pages 536-540, 1995.

[7] P. Daras et al., Three-dimensional shape-structure comparison method for protein classification, *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 3(3), pages 193-207, 2006.