UCRL-JRNL-209388

Computational approaches for identification of conserved/unique binding pockets in the A chain of ricin

Carol L. Ecale Zhou, Adam T. Zemla, Diana Roe, Malin Young, Marisa Lam, Joe Schoeniger, Rod Balhorn

February 2, 2005

Bioinformatics

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Please address correspondence regarding this manuscript to:

Carol L. Ecale Zhou
Pathogen Bioinformatics
Energy, Environment, and Biology Division
L-174
Lawrence Livermore National Laboratory
7000 E. Avenue
Livermore, CA 94550 USA
Phone: 925-422-2117
Fax: 925-423-6437
e-mail: zhou4@llnl.gov

For submission to *Bioinformatics*

Computational approaches for identification of conserved/unique binding pockets in the A chain of ricin

Carol L. Ecale Zhou[1], Adam T. Zemla[1], Diana Roe[2], MalinYoung[2], Marisa Lam[1], Joseph S. Schoeniger[2], and Rod Balhorn[1]

[1]Lawrence Livermore National Laboratory, [2]Sandia National Laboratory

UCRL-JRNL-209388

BIOINF-2005-0307

STRUCTURED ABSTRACT

Motivation: Specific and sensitive ligand-based protein detection assays that employ antibodies or small molecules such as peptides, aptamers, or other small molecules require that the corresponding surface region of the protein be accessible and that there be minimal cross-reactivity with non-target proteins. To reduce the time and cost of laboratory screening efforts for diagnostic reagents, we developed new methods for evaluating and selecting protein surface regions for ligand targeting.

Results: We devised combined structure- and sequence-based methods for identifying 3D epitopes and binding pockets on the surface of the A chain of ricin that are conserved with respect to a set of ricin A chains and unique with respect to other proteins. We 1) used structure alignment software to detect structural deviations and extracted from this analysis the residue-residue correspondence, 2) devised a method to compare corresponding residues across sets of ricin structures and structures of closely related proteins, 3) devised a sequence-based approach to determine residue infrequency in local sequence context, and 4) modified a pocket-finding algorithm to identify surface crevices in close proximity to residues determined to be conserved/unique based on our structure- and sequence-based methods. In applying this combined informatics approach to ricin A we identified a conserved/unique pocket in close proximity (but not overlapping) the active site that is suitable for bi-dentate ligand development. These methods are generally applicable to identification of surface epitopes and binding pockets for development of diagnostic reagents, therapeutics, and vaccines.

Contact: zhou4@llnl.gov

INTRODUCTION

Ricin is a potent toxin of plant origin and is well known for its use as a biological weapon as well as for in vivo and in vitro therapeutic applications (Day et al. 1996, Knight 1979, Lord et al. 1994, Olsnes and Kozlov 2001). The holotoxin consists of co-translated A and B chains, which are post-translationally cleaved but covalently rejoined by a disulfide linkage (Lord et al. 1994). The B chain mediates transport of the A chain to the cytoplasm where the A chain inactivates ribosomes by depurinating 28S RNA (Lord et al. 1994, Wesche et al. 1999). Ricin shares extensive sequence homology with lectins from a wide range of organisms.

Our interest in ricin is motivated by the development of high-affinity reagents that will selectively identify ricin for applications in bio-defense. Due to the great increase in genomic data in recent years, it has become possible to apply informatics approaches to the identification of molecular targets for highly specific and selective identification of pathogens (Slezak et al. 2003, Gardner et al. 2004). Whereas DNA-based reagents can be developed using any portion of a genome and application of linear hybridization-based technologies, antibody-based assays require that the proteins they target be expressed in reasonable abundance and usually that the corresponding surface regions be accessible on the intact target protein. If synthetic ligands such as peptides, aptamers, or other small molecules are used in a molecular recognition/affinity-based assay, the binding site must likewise be accessible and in reasonable abundance for the assay to function.

The affinity of a ligand for its protein target is far less predictable than are the experimental conditions that will distinguish among nearly identical sequences in a hybridization reaction. In addition to the requirements for abundance and accessibility of the target binding site, we must also attempt to predict which epitopes and binding pockets will generate low cross reactivity and good affinity. Due to the complexities of protein structure, it is difficult to predict which proteins might pose a risk for cross reactivity. This set of requirements projects substantial challenges relative to the design of DNA hybridization assays, where abundance and accessibility are presumed to be uniform across the genome, and the tradeoffs of affinity with sequence uniqueness and cross reactivity can be quantitatively predicted.

Purely informatics methods for comparing sequences and structures have proven useful in identifying regions of structural or functional significance in proteins (Fygenson et al. 2004, Karlin and Ghandour 1985, Ouzounis et al. 2003, Zemla 2003). Reagents commonly used for protein detection include monoclonal and polyclonal antibodies (Peruski and Peruski 2003), small molecule ligands (Lightstone et al. 2000), and natural or synthetic peptides (Wang et al 2004). However, targets for these reagents have not typically been identified using informatics approaches. In analyzing ricin A, we devised a combined informatics approach that allowed us to identify a region on the protein surface that distinguished ricin A from other proteins. We used the Protein Data Bank (PDB) (Berman et al. 2000) and the National Center for Biotechnology Information's non-redundant (NR) protein sequence databases (http://www.rcsb.org/pdb; http://www.ncbi.nlm.nih.gov/) and combined methods of structure comparison and sequence-based analyses to identify residues comprising subsequences or surface regions that were conserved with respect to ricin A, yet unique with respect to other proteins. We propose here a set of informatics tools for identifying optimal binding pockets and epitopes for targeting protein-based detection reagents. In this paper we describe the methods by which we were able to identify a unique structural feature in ricin A that could be targeted for developing reagents for small-molecule or antibody binding assays. These methods are broadly applicable to identification of structural features or binding pockets for development of diagnostics, therapeutics, and vaccines.

SYSTEM AND METHODS

Structural comparison of ricin and near-neighbor proteins.

As a reference sequence for our analyses of the A chain of ricin, we chose an entry (ID=RICI_RICCO, P02879) from the SHIGARICIN family of the PRINTS database of virulence factors (Attwood et al. 1997, http://www.jenner.ac.uk/BacBix3/Ppprints.htm). Of the 21 PDB structures (sets of experimentally solved atom coordinates) of the ricin A chain, we chose the three non-redundant, non-mutant structures that had been solved with highest resolution (PDB entries 1br6, 1br5, and 1rz0;Yan et al. 1997, Gabdoulkhakov et al. to be published) to include in the target set for our structure-based analyses. These structures had sequence similarity of between 93% and 100% (and corresponding structure similarity LGA_S score between 95% and 100%) to the ricin A reference. Using the AS2TS (Zemla et al. in press) automated homology based protein structure modeling system, PDB entry 1br6 was selected as the 3D model structure of the ricin A reference sequence because it had the greatest sequence similarity (100% sequence identity) and structure completeness from among available PDB structures (100% of structure solved with resolution of 2.3 Å). The 3D model of the reference ricin A chain was then used to assemble a complete list of all related structures from PDB using LGA (local-global alignment) software, which employs a PDB search method (Zemla 2003). In our PDB search for related structures, those representing mutants or redundant sequences were removed from consideration; we selected 31 structures from PDB that had significant similarity to our reference model. LGA was used to structurally align the reference structure with the other ricin structures (targets) and with all other related structures (near neighbors) (sequence similarity between 13% and 40%, and corresponding LGA_S between 59% and 87%). The reference model of ricin A was structurally superimposed and compared with all other structures, and the results of the comparison were sorted by structure similarity score (Fig. 1). Our PDB search, along with the observed clear structural distinction between ricins (LGA_S above 93%) and 31 close homologs (near neighbors, LGA_S above 59%) allowed us to define a complete set of 34 ricin A or ricin A-like proteins from PDB. We observed a clear structural distinction between the 34 ricin A and related structures and all other structures from PDB; the closest ones had a level of structure similarity no higher than 20%. The ricin A-like proteins mostly consisted of plant lectins with ribosome-inhibiting activity. For each pair-wise alignment using LGA, distances between corresponding C-alpha carbons were computed. The structure comparison plots (Fig. 1) were examined for regions of structure similarity among the ricin A structures and structure deviation between ricin A and near neighbors.

cuScore residue correspondence analysis.

The cuScore was devised as a measure of residue conservation/uniqueness in structure context. The sequence homology between ricin and a large number of other proteins (e.g., plant lectins), and the structural homology among ricin and lectins from widely ranging taxa, prompted us to consider the challenge of identifying binding pockets and epitopes for ricin A diagnostic reagents that would pose minimal risk of cross reactivity from related proteins. Because structure is more highly conserved than is sequence, we determined that sequence-based analysis among ricin A homologs may be insufficient. Therefore, we developed the cuScore as a means of comparing corresponding residues across a set of structural homologs.

A multiple structure-based residue-residue correspondence, which resembles a multiple-sequence alignment, was extracted from the LGA comparison (Fig. 1), and corresponding structurally aligned residues in targets vs. near-neighbor homologs were compared. For each corresponding (co-aligned) position in the set of target proteins, a consensus amino acid was identified as the residue that occurred in more than half of the members of the set. A conservation measure, c, per position was computed between 0 and 1 depending on the degree of conservation at a given position among the corresponding residues of the target set. For each corresponding residue in the set, a score of 1 was assigned if the residue matched the consensus, 0.5 if the residue did not match but occurred within the same amino-acid group ((AGILPV), (FWY), (DE), (RKH), (ST), (CM), (NQ)), or 0 if the residue occurred within a different group. Thus, scores of 1, 0.5, and 0 represented decision states of identity, similarity, and dissimilarity, respectively. Our choice of amino-acid groupings for cuScore analysis was based on a grouping that we felt was most appropriate for identifying binding pockets or epitopes suitable for our purposes—namely, one that grouped amino acids based on chemistry and size: aliphatic (AGILPV), aromatic (FWY), acidic (DE), basic (RKH), small hydroxylic (ST), sulfur-containing (CM), and amidic (NQ). Although other grouping schemes (amino-acid alphabets) based on physical properties, substitution propensity, codon degeneracy, or kinetic properties (Karlin and Ghandour 1985, Fygenson et al. 2004) may have yielded quantitatively different results, we did not consider these alphabets to be any more appropriate for our goal of identifying conserved/unique regions for recognition by means of ligand binding. The conservation measure (c) was computed as the sum of the scores divided by the number of residues in the set, thereby representing an average degree of similarity among corresponding residues. Residues that had not been assigned spatial coordinates were not included in the set. If a consensus residue could not be identified (indicating poor conservation at that position within the target set), the cuScore was immediately set to 'undefined' with no further calculation. Each structurally corresponding residue in the set of near-neighbors was then compared to the consensus residue to determine a uniqueness measure (u). Near-neighbor residues were scored according to similarity by determining whether they were identical (score = 1), were different but occurred within the same amino-acid group (score = 0.5), or occurred in a different group (score = 0). Positions representing deletions with respect to the reference ('-' in the alignment) were scored 0, as were positions that had been labeled 'X' (e.g., non-standard amino acid) in the coordinates file. The uniqueness measure (u) was calculated as the sum of the scores divided by 31. For each residue in the reference ricin sequence, a cuScore was computed as c - u. In the case that a cuScore were computed to be less than 0, (indicating greater conservation at that position among near neighbor proteins than within the targets), cuScore would be set to 'undefined'. cuScores were plotted vs. ricin A residue number (Fig. 2). cuScores for residues at the extreme N- and C-terminal regions, for which there was inadequate structural data, were also left 'undefined'.

pScore subsequence analysis.

The pScore was devised as a measure of residue infrequency in local sequence context. We applied a "sliding window" approach to generate all subsequences of length n (n = 4, 5, or 6) for the reference ricin A sequence. Initial tests using larger window sizes (up to 10) indicated that window sizes above 6 yielded matches primarily to ricins and close homologs and, therefore, did not provide useful information about potential cross-reactivity posed by distantly related or unrelated sequences. We then determined how often each subsequence occurred in the NR database. This was called a subsequence's "popularity". For each residue in the reference sequence, a score was computed as the sum of the popularity values for each of the n windows that the residue was a member of, divided by n (= average popularity for a set of n windows containing the residue), thus each window was weighted equally. We saw no justification for assigning greater weight to any window within a given pScore calculation; we felt this was justified because residue side-chains tend to alternate direction in 3-D space, and therefore a set of residue participating in a ligand binding reaction may tend to occupy alternating positions along the chain. Using pScore analysis, we were looking for regions that occurred with relative infrequency. For our convenience, we normalized the scores so that they would range from 0 to 1 and could be meaningfully plotted alongside cuScores. Normalization was done as follows. The minimum and maximum scores were determined (from among scores of all subsequences) and pScore was computed as 1 - (score – minimum)/(maximum – minimum). pScores for residues at the N- and C-termini for which there was incomplete data (fewer than n windows containing a given residue) were left undefined. pScores were plotted against ricin A residue number alongside cuScores (Fig. 2).

Determination of surface residues with high cuScores and pScores.

Visualization of surface-exposed regions containing residues with high cuScores and/or pScores was facilitated using Rasmol (Sayle and Milner-White 1995) to color code low- to high-scoring residues (Fig 3). cuScores and pScores were separately loaded into the b-factor column of the reference ricin A 3D coordinates file and displayed using Rasmol's color-temperature setting. We used cuScore and pScore values and these 3D color plots, along with naccess (Hubbard and Thornton 1993) solvent accessibility calculations (data not shown), to visually identify surface loops or binding pockets suitable as antibody or small molecule ligand targets. By visual inspection of the residues comprising region R2 (Fig. 1), we determined that this subsequence was composed mostly of residues with high cuScores and pScores (Figs. 2, 3).

Use of UniquePocket software to identify crevices for small-ligand binding.

UniquePocket is a tool for automating and objectifying the process of identifying ligand-binding crevices (pockets) that occur in close proximity to residues of interest (e.g., conserved in the protein of interest but unique with respect to near neighbor proteins). Ideally, pockets should be close to several conserved/unique residues (hereafter referred to as "unique"), which may not necessarily be adjacent to each other in sequence space, but are in close proximity in 3-D space. We created the UniquePocket program to identify these pockets, using site-identification tools from the DOCK suite (Kuntz et al. 1982). UniquePocket has 3 steps: 1) identification of all protein crevices by filling them with spheres to represent the negative image of the protein; 2) labeling each sphere as 'unique' or 'not unique'; 3) clustering the unique spheres to generate suggested 'unique pockets'. UniquePocket uses the SPHGEN program from DOCK to create the initial spheres and the Cluster program to perform the final clustering. For labeling spheres, we used the uniqueness information from the occupancy column, with cutoffs set so a unique sphere had to occur within 8Å distance from 30 labeled atoms (for an average interaction with 3 labeled residues). The output of UniquePocket is 3 files: unique.sph, cluster.sph, and unique_cluster.sph. unique.sph contains the full set of all unique spheres and represents the "unique" volume of the protein. The two files, unique.sph and cluster.sph, are different types of sphere clusters or "pockets" made from only unique spheres and unique spheres combined with regular spheres, respectively. We used a cutoff minimum of 5 unique spheres per cluster for the second set of spheres. The above cutoffs were selected so that a single unique region could produce a single, small cluster, but regions with overlapping unique regions would be favored and produce several larger clusters of spheres. We did a study to determine the sensitivity of this approach with respect to slight structural deviations (Table 1). We ran UniquePocket on several sets of crystal structures with different resolution, and also homology models of proteins in which side chain placements were calculated using SCRWL (Bower et al 1997, Canutescu et al 2003). As seen in Table 1, the exact cluster boundaries were very sensitive (number ranging from 7-9) to protein shape and sidechain position, although the overall unique volume was fairly (volume varying by <10%) robust.

IMPLEMENTATION

Structure comparison between ricin A and ricin A-like proteins.

Inspection of LGA structure comparisons of three ricin A structures and 31 ricin A-like (near-neighbor) structures yielded 6 regions of structural interest (Fig. 1). Regions R1, R2, and R5 showed structural conservation among the ricin A structures and deviation of at least 6 Å Ca-Ca distance between corresponding residues of all near-neighbor proteins. Region R6 also showed structural conservation among the ricin A structures, although that conservation was also seen in nine of the highest-scoring (LGA_S = 85.993 and above) near neighbors. Regions R3 and R4 showed inconsistency in structural deviation among near neighbors in that several low-scoring proteins appeared to be well aligned, and R4 showed inconsistent structural conservation among the ricin A structures. Based on consistency with respect to structural conservation among ricin A chains and deviation among near-neighbor structures as compared to ricin A, regions R1, R2, R5, and R6 were selected as regions of interest for purposes of optimal epitope and binding pocket discovery. Our PDB search, along with the observed clear distinction between 31 close homologs and the more distant structures (of which only 3 are shown at bottom of graph) allowed us to define a complete set of 34 non-redundant ricin A or ricin A-like proteins from PDB. Regions R1-R6 were transposed into subsequence residue numbering to correspond with the ricin A reference sequence as follows: R1 (48-54), R2 (93-100), R3 (117-120), R4 (129-134), R5 (151-158), R6 (192-195). To examine the possibility that detection of structural deviations using this method could be confounded by experimental (e.g., crystallographic) margins of error or by expected deviations in regions that are by nature disordered, we used LGA to superimpose all 21 ricin A structures from PDB (data not shown). Aside from significant (>3Å) deviations identified at the N- and C-termini, we found small (1Å or less), consistent deviations spanning short subsequences (1-6 residues) only within region R5. The other regions of interest each displayed only deviations of 1Å or less in fewer than half of the 21 structures.

cuScore and pScore analyses.

cuScores and pScores for ricin A were plotted against residue number (Fig. 2). Superimposed on this plot were regions R1 through R6 determined by LGA structure comparison (Fig. 1) and subsequences corresponding to PRINTS fingerprints for the SHIGARICIN family (http://www.jenner.ac.uk/BacBix3/PPprints.htm). Region R2 was determined to be composed mostly of residues with high cuScores (F93, 0.32, H94, 0.81; P95, 0.94; D96, 0.98; N97, 1.00; Q98, 0.98; E99, 1.00; D100, 0.98) and high pScores (data for window size 4: F93, 0.96, H94, 0.96; P95, 0.95; D96, 0.95; N97, 0.94; Q98, 0.92; E99, 0.84; D100, 0.74). pScore plots using window sizes of 4, 5, 6, and 7 were compared. We found that a pScore window size of 7 generated matches to NR sequences comprising primarily ricin and ricin-like sequences, and therefore was less useful to our goal of quantifying residue infrequency for the purpose of excluding from consideration regions on ricin A that would pose risk of reagent cross-reactivity from distantly related proteins. pScore results for window sizes 4, 5, and 6 were not qualitatively different (data not shown), although the distribution of scores (curve) was compressed toward 1 with increasing window size. For ease of illustration, therefore, we report pScores at window size 4 (Fig. 2).

Identification of crevise suitable for ligand binding to uniquely identify ricin A.

We applied the UniquePocket program to identify the most suitable crevices in ricin A for subsequent development of reagents (small chemical ligands) for specific identification of ricin A. As input to UniquePocket we used the reference ricin A 3D coordinates file, modified so as to contain data in the "occupancy" column to mark those residues corresponding to regions R1, R2, R5, and R6. We combined the top clusters from each clustering file to create the final sphere cluster shown in Fig. 4. This site is not only unique, but also in close proximity (8Å centroid difference) to the sugar-binding site in ricin A. The crevice identified by UniquePocket analysis was beneath region R2 and adjacent to region R1 (Fig. 4). Region R2 contained residues that were both conserved/unique based on cuScore analysis and infrequent in local sequence context based on pScore analysis (Figs. 2, 3).

Our sensitivity experiments with UniquePocket showed that this technology is sufficiently robust against small structural deviations when identifying conserved/unique binding pockets (Table 1). There was no appreciable difference in the characteristics of the crevice identified by UniquePocket analysis when we compared two crystal structures of ricin A solved with different resolutions (two first rows in Table 1), with two homology models (different sidechain placements calculated using SCWRL), and a medium homology model based on a template with 40% sequence identity. Our technology was not, therefore, dependent on the availability of a solved crystal structure, and the homology 3D model created using our AS2TS modeling system yielded qualitatively similar results (Table 1).

Validation of pScore, cuScore, and UniquePocket predictions.

To test whether the pockets we predicted to be unique would actually generate unique (predicted) docking results, we compared docking scores to the R2 unique site, and to the active site, which we predicted would not be unique. We used DOCK4 (Ewing et al. 2001) to dock 1000 randomly selected compounds to both these sites within a selection of ricin and near-neighbor proteins. We chose 3 ricin structures (1br5, 1br6, and 1rzo) and 3 close near neighbors (1abr, 1mrj, and 1tfm), each with at least 38% homology to the ricin reference sequence (Table 2.) Among the ricins, we found high correlations in docking score at both the active site and the R2 conserved/unique site, as expected. Among the near-neighbor proteins, we found high correlations in docking score at the active site, and some correlation at the active site between ricins and near neighbors. At the unique site, we found some correlation among the near neighbors and very little correlation between the ricins and the near neighbors.

DISCUSSION

Structure comparison of ricin A with ricin A-like (near-neighbor) proteins, and mapping of cuScores and pScores onto a 3D model of ricin A (Fig. 3) enabled us to identify surface regions that are suitable sites for directing development of detection reagents (e.g., small molecule ligands or antibodies). Application of our UniquePocket software identified a crevice (Fig. 4) that is defined by a region (R2) of structural conservation/uniqueness (conserved within ricin A structures and unique with respect to near-neighbor proteins) (Figs 1, 2), which also coincides with residues determined by cuScore to be conserved/unique and determined by pScore analysis to be infrequent in local sequence context (Figs. 2-4). These residues are also in close proximity to, but do not overlap, the active site of ricin A (Figs. 3,4). This combined informatics approach identified a binding pocket on the A chain of ricin with the desired qualities for development of a bi-dentate ligand that would be expected to identify ricin A based on its biological activity and to distinguish it from other proteins. A bi-dentate ligand consists of two small molecules covalently bonded to a linker of appropriate length to allow binding of each small molecule to its respective binding site (Lightstone et al. 2000, Shuker et al 1996, Wang et al 2004). The affinity of a bi-dentate ligand has been shown to increase greatly when compared to the affinities of the two individual ligands.

Targeting the ricin A active site alone would likely be insufficient for distinguishing ricin A from other proteins with ribosome-inactivating activity, given that the residues that make up the active site of ricin A (Lord et al. 1994, Marsden et al. 2004, Weston et al 1994) showed low cuScores, indicating a low degree of conservation/uniqueness (Y80, 0.03; Y123, 0.00; E177, 0.00; R180, 0.00; W211, 0.00) (Fig. 3). Residues defining the surface loop within region R2, however, showed high cuScores and pScores, as well as structural deviation as determined by LGA structure alignment. Additionally, our docking test (Table 2) demonstrated differences in ligand docking between ricins and structurally related proteins at the unique site, but much less so at the active site. We believe that the crevice lying beneath this loop would be, therefore, the most suitable binding site to target for generating a reagent that could distinguish ricin A from structurally similar proteins and from dissimilar proteins that have surface-exposed regions composed of similar subsequences. In addition, this surface exposed subsequence could be used for development of peptide-based antibodies, which could then be affinity matured against the whole protein.

Previous work (Lebeda and Olson 1999, Olson et al. 2004) used informatics methods to examine the physical and chemical properties of a 26-residue loop-helix-loop structure of ricin A (residues Y91-T116) that had been shown to elicit a strong and specific antibody response. This loop-helix-loop contains region R2 (residues 93-100). We note that the structure alignment that we obtained using LGA differs from that reported previously in that a gap introduced in our alignment to accommodate shorter loops in near-neighbor proteins occurs in the vicinity of residues 97-98 (alignment data not shown; see Fig. 5 for structure comparison), whereas Lebeda and Olson's (1999) alignment introduced a gap at residues 106-107. Residue D96, which has been hypothesized to be conserved in all ribosome-inhibiting proteins (alternating with glutamate), yielded a cuScore of 0.98. This high score is accounted for by our structure alignment, which is based on optimization of C-alpha-C-alpha deviation (Zemla 2003).

However, close examination of the alignment suggests that shifting a single residue to the right within the near-neighbor sequences would align the ricin A chain's D96 with aspartate and glutamate residues of the near neighbor sequences. Such a shift was not justified based on the C-alpha-C-alpha structure alignment, although a structure alignment based on beta carbons did indeed align D96 with most of the corresponding D or E residues in the near-neighbor proteins (data not shown). This observed conservation of a negatively charged residue believed to be involved in the binding of the rRNA substrate (Lebeda and Olson 1999) suggests that any sequence- or structure-based alignment is subject to interpretation based on functional annotation. The difficulty of establishing a true and consistent alignment does not, however, diminish the relevance of a purely informatics approach for predicting surface regions of interest in developing reagents that have a high degree of specificity. The resulting high cuScore of residue D96 is a consequence of the structural deviation within this variable loop. By quantifying structure and sequence deviation among target (ricin) and near-neighbor (other RIP) proteins, our methods detected region R2 and suggest that this loop may contain residues of greatest interest in defining antibodies that distinguish ricin A from similar proteins.

It is interesting to note that regions of structural deviation determined by LGA analysis (R1-R6, Figs. 1, 2) have little overlap with SHIGARICIN fingerprints assigned to ricin A (http://www.jenner.ac.uk/BacBix3/Ppprints.htm). Only R4 partly overlaps with the SHIGARICIN3 fingerprint. This is not surprising in that fingerprint motifs that characterize a family of proteins would be expected to represent structurally conserved regions, whereas our LGA analysis identified regions of backbone deviation between ricin proteins and other structurally similar proteins.

Whereas plotting cuScores and pScores allowed us to identify subsequences comprising conserved/unique and "infrequent" residues (Fig. 2), mapping of these measures onto the structure model (Fig. 3) allowed us to also visualize in 3-D space surface regions formed by folding together non-contiguous residues. This capability is widely applicable for identification of regions suitable for development of diagnostic reagents for proteins expressed by pathogens or associated with disease, virulence, toxicity, or for development of therapeutic drugs or antibodies, and may reduce the time and cost of such efforts by identifying up front those regions that are optimal for reagent targeting in terms of specificity for the proteins of interest and that pose the least risk in terms of cross reactivity by other proteins. Furthermore, a 3D mapping of surface residues identified by these methods could be helpful in predicting the specificity of antibody reagents that have been epitope mapped to their target protein. Experimental validation of this informatics approach would entail comparing the specificities (rates of false positives and negatives) of antibody and high-affinity ligand reagents that recognize surface regions displaying high vs. low cuScores and pScores.

REFERENCES

Attwood,T.K., Avison,H., Beck,M.E., Bewley,M., Bleasby,A.J., Brewster,F., Cooper,P., Degtyarendko,K., Geddes,A.J., Flower,D.R., Kelly,M.P., Lott,S., Measures,K.M., Parry-Smith,D.J., Perkins,D.N., Scordis,P., Scott,D., and Worledge,C. (1997) The PRINTS database of protein fingerprints: A novel information resource for computational molecular biology. J Chem Inf Comput Sci, **37**, 417-424.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., and Bourne,P.E. (2000) The protein data bank. Nucleic Acids Research, **8**, 235-242.

Bower,M.J., Cohen,F.E. and Dunbrack,R.L. (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol, **267**, 1268-1282.

CanutescuA.A., Shelenkov,A.A. and Dunbrack,R.L. (2003) A graph theory algorithm for protein side-chain prediction. Prot Sci, **12** ,2001-2014.

Day,P.J., Ernst,S.R., Frankel,A.E., Monzingo,A.F., Pascal,J.M., Molina-Svinth,M.C. and Robertus,J.D. (1996) Structure and activity of an active site substitution of ricin A chain. Biochemistry, **35**, 11098-11103.

Ewing, T.J.A., S. Makino, A.G. Skillman, I.D. Kuntz. 2001. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. Journal of Computer-Aided Molecular Design 15: 411-428.

Fygenson,D.K., Needlemen,D.J. and Sneppen,K. (2004) Variability-based sequence alignment identifies residues responsible for functional differences in a and b tubulin. Protein Science, **13**, 25-31.

Gabdoulkhakov,A.G., Savochkina,Y., Konareva,N., Krauspenhaar,R., Stoeva,S., Nikonov,S.V., Voelter,W., Betzel,C., Mickhailov,A.M.. Structure-Function Investigation Comlex of Agglutinin from Ricinus Communis with Galactoaza (to be published).

Gardner,S., Lam,M.W., Mulakken,N.J., Torres,C.L., Smith,J.R. and Slezak,T.R. (2004) Sequencing needs for viral diagnostics. Journal of Clinical Microbiology, **42**, 0095-1137.

Hubbard,S.J. and Thornton,J.M. (1993) ‘NACCESS’, Computer Program, Department of Biochemistry and Molecular Biology, University College, London.

Karlin,S. and Ghandour,G. (1985) Multiple-alphabet amino acid sequence comparison of the immunoglobulin k-chain constant domain. Proc. Natl. Acad. Sci. USA, **82**, 8597-8601.

Knight,B. (1979) Ricin—a potent homicidal poison. British Medical Journal, **278**, 350-351.

Kuntz,I.D., Blaney,J.M., Oatley,S.J., Langridge,R. and Ferrin,T.E. (1982) A geometric approach to macromolecule-ligand interactions. J. Mol. Biol., **161**, 269-288.

Lebeda,F.J. and Olson,M.A. (1999) Prediction of a conserved, neutralizing epitope in ribosome-inactivating proteins. International Journal of Biological Macromolecules, **24**, 19-26.

Lightstone,F.C., Prieto,M.C., Singh,A.K., Piqueras,M.C., Whittal,R.M., Knapp,M.S., Balhorn,R. and Roe,D.C. (2000) Identification of novel small molecule ligands that bind to tetanus toxin. Chem Res Toxicol., **13**, 356-362.

Lord,J..M., Roberts,L.M. and Robertus,J.D. (1994) Ricin: structure, mode of action, and some current applications. FASEB J, **8**, 201-208.

Marsden,C.J., Fulop,V., Day,P.J and Lord,J.M. (2004) The effects of mutations surrounding and within the active site on the catalytic activity of ricin A chain. Eur. J. Biochem., **271**, 153-162.

Olson,M.A., Carra,J.H., Roxas-Duncan,V., Wannemacher,R.W., Smith,L.A., and Millard,C.B. (2004) Finding a new vaccine in the ricin protein fold. Protein Engineering, Design & Selection, **17**, 391-397.

Olsnes,S. and Kozlov,J.V. (2001) Ricin. Toxicon 39:1723-1728.

Ouzounis,C.A., Coulson,R.M., Enright,A.J., Kunin,V., Pereira-Leal,J.B. (2003) Classification schemes for protein structure and function. Nat Rev Genet., **4**, 508-519.

Peruski,A.H., and Peruski,Jr,L.F.. (2003) Immunological methods for detection and identification of infectious disease and biological warfare agents. Clinical and Diagnostic Laboratory Immunology, **10**, 506-513.

Portefaix, J.-M., S. Thebault, F. Bourgain-Guglielmetti, M.D. Del Rio, C. Granier, J.-C. Mani, I. Navarro-Teulon, M. Nicolas, T. Soussi, and B. Pau. 2000. Critical residues of epitopes recognized by several anti-p53 monoclonal antibodies correspond to key residues of p53 involved in interactions with the mdm2 protein. Journal of Immunological methods 244: 17-28.

Sayle,R.A. and Milner-White,E.J.. 1995. RasMol: Biomolecular graphics for all. Trends in Biochemical Sciences, **20**, 374-376.

Shuker,S.B., Hajduk,P.J., Meadows,R.P. and Fesik,S.W. (1996) Discovering High-Affinity Ligands for Proteins: SAR by NMR. Science, **274**, 1531-1534.

Slezak,T., Kuczmarski,T., Ott,L., Torres,C., Medeiros,D., Smith, J., Truitt,B., Mulakken,N., Lam,M., Vitalis,E., Zemla,A., Zhou,C. E. and Gardner,S. (2003) Comparative genomics tools applied to bioterrorism defense. Briefings in Bioinformatics, **4**, 133-149.

Wang,G., De,J., Schoeniger,J.S., Roe,D.C. and Carbonell,R.G. (2004) A hexamer peptide ligand that binds selectively to staphylococcal enterotoxin B: isolation from a solid phase combinatorial library. Journal of Peptide Research, **64**, 51-64.

Wesche,J., Rapak,A. and Olsnes,S. (1999) Dependence of ricin toxicity on translocation of the toxin A-chain from the endoplasmic reticulum to the cytosol. J Biol Chem, **274**, 34443-34449.

Weston,S.A., Tucker,A.D., Thatcher,D.R., Derbyshire,D.J. and Pauptit,R.A. (1994) X-ray structure of recombinant ricin A-chain at 1.8 Å resolution. J. Mol Biol., **244**, 410-422.

Yan,X., Hollis,T., Svinth,M., Day,P., Monzingo,A.F., Milne,G.W., Robertus,J.D. (1997) Structure-based identification of a ricin inhibitor. J Mol Biol, 266, 1043.

Zemla,A. (2003) LGA: a method for finding 3D similarities in protein structures. Nucleic Acid Research, **31**, 3370-3374.

Zemla,A., Ecale Zhou,C., Slezak,T., Kuczmarski,T., Rama,D., Torres,C, Sawicka,D. and Barsky,D. (2005) AS2TS system for protein structure modeling and analysis. Nucleic Acids Research, in press (web access: http://as2ts.llnl.gov/)

ACKNOWLEDGEMENTS

This work was performed under the auspices of the U. S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48 and was supported by funding from the Department of Homeland Security.

FIGURES

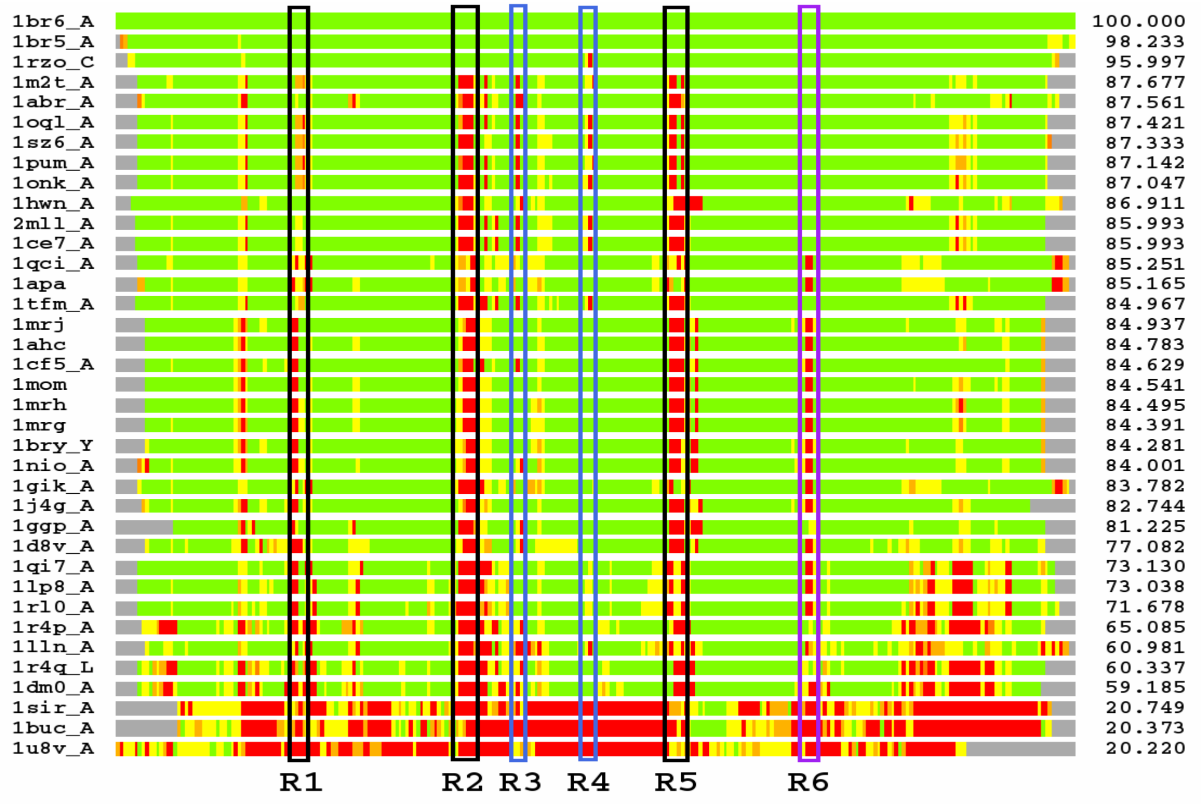

Fig 1. LGA pair-wise structural comparison of ricins, ricin-like proteins, and distant structural homologs. Bars: proteins (N-terminus at left, C-terminus at right) that have been aligned with the reference ricin A structure. Bars are shown sorted by LGA_S in descending order. Ricin A PDB structures comprise top 3 bars, ricin-like homologs comprise the next 31 bars, and the last 3 bars represent distant structure homologs. Left column: PDB identifier; right column: LGA_S measure of structure similarity. Colors represent distance deviations between C-alpha carbons: green <2.0 Å, orange 2.0-<4.0 Å, yellow 4.0-<6.0 Å, red >=6.0 Å. Boxes (R1-R6) delineate regions of structural conservation among ricin A structures and deviation from near-neighbors.

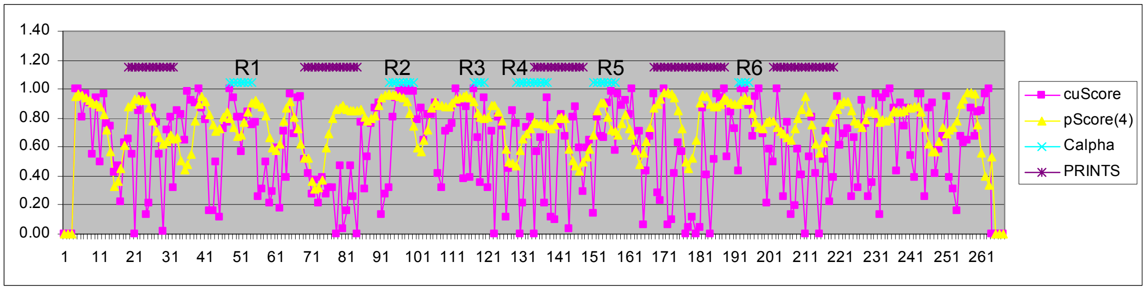

Fig 2. Plots of cuScore and pScore values for residues of ricin A. (Pink squares) cuScores and (yellow triangles) pScores at window size = 4. R1-R6: regions of structural conservation/uniqueness determined using structure alignment analysis (see Fig. 1). SHIGARICIN fingerprints are shown as dark red horizontal segments above regions R1-R6. Depicted are pScores for window size = 4.

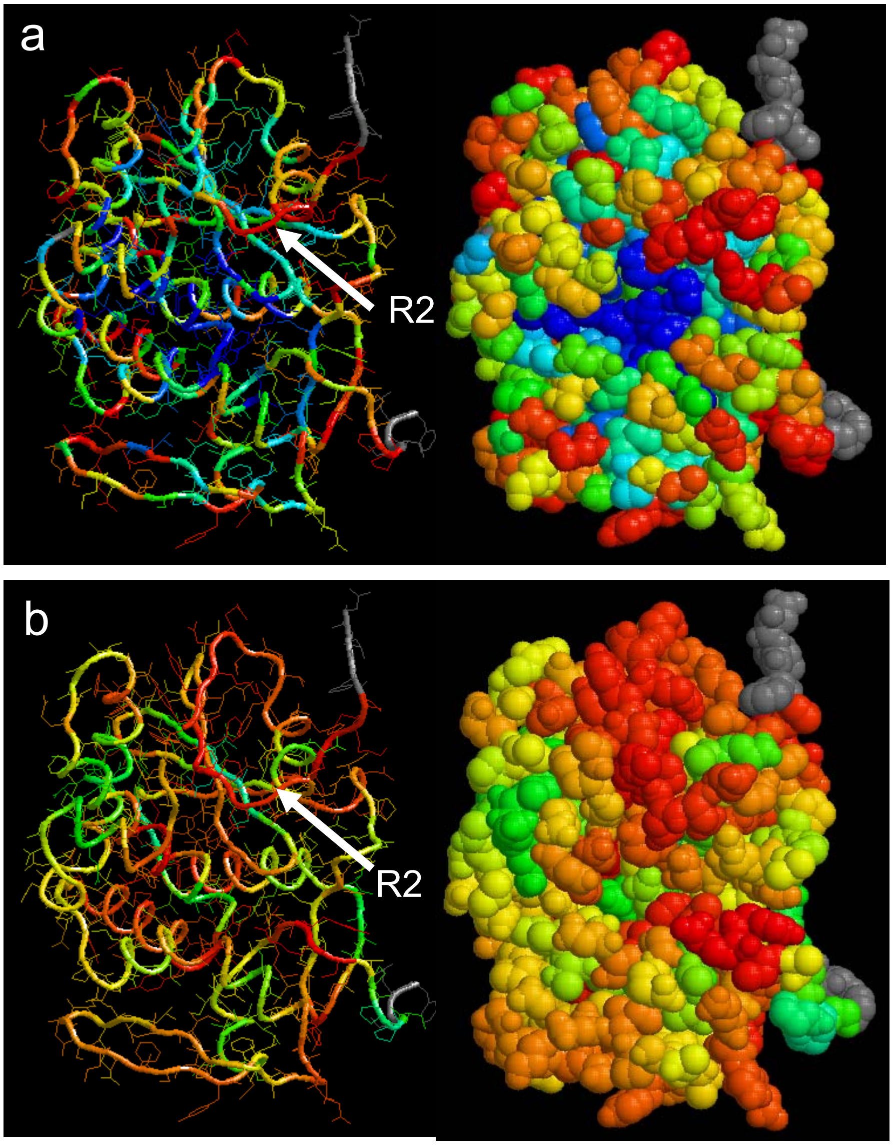

Fig 3. Mapping of a) cuScores and b) pScores to structure of ricin A using B-factor column and temperature factor color setting in Rasmol. Wire frame and space fill views are shown for each. Values range from low to high (range 0.0 to 1.0, see Fig 2.) as: blue – green – yellow – orange – red. Grey indicates undefined scores at the N- and C-terminal regions. Arrow points to a central residue within region R2 (see Figs. 1, 3). Region R2 (arrows in a, b) contains residues with high cuScores (a) and high pScores (b). The blue residues forming a pocket (center of space fill image in panel a) comprise the active site.

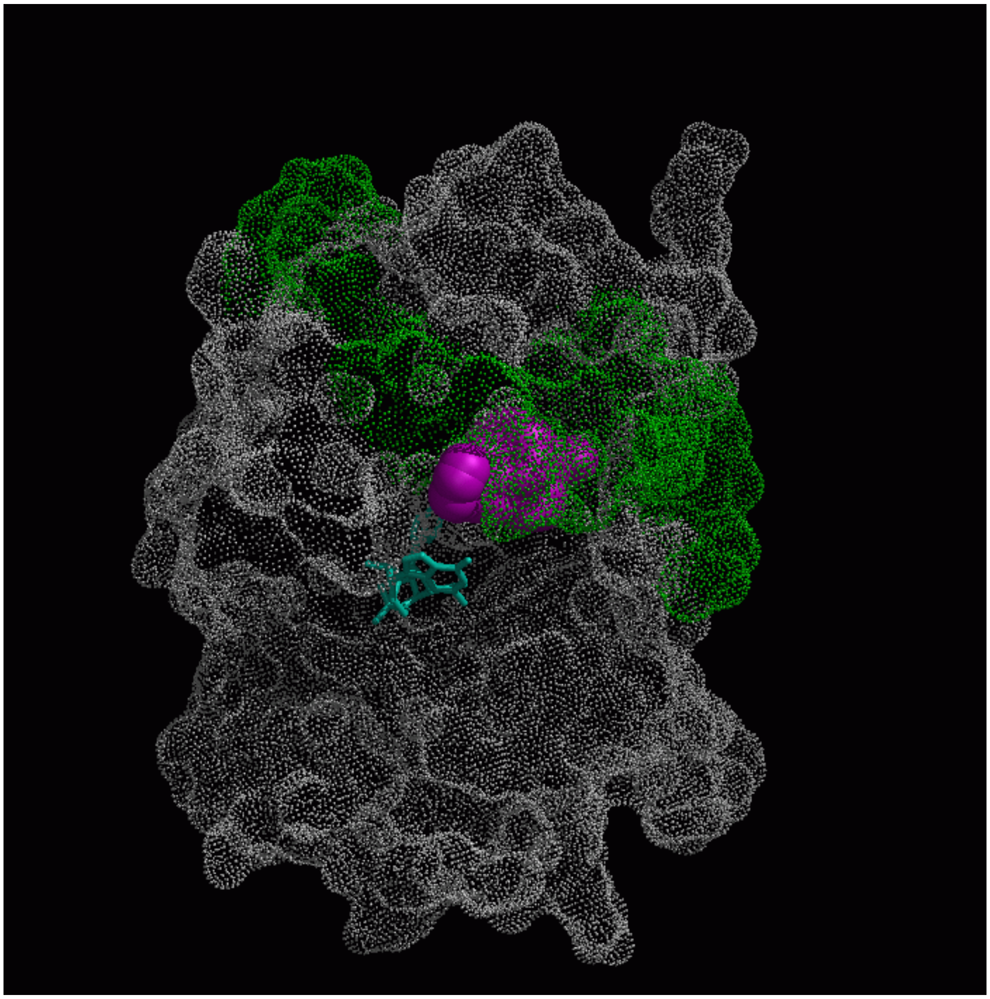

Fig. 4. Ricin A structure showing crevice identified by UniquePocket software. Magenta: spheres filling volume that determines crevice suitable for small-ligand binding. Green: Regions R1, R2, R5, and R6 (see Fig. 1). Cyan: sugar substrate analog bound to active site.

Table 1. Definition of UniquePocket is robust against small structural deviations: comparing proteins and homology 3D models with different sidechain positions.

Template for homology structure	Percent sequence homology to ricin	Num. of unique sphere clusters	Num.of unique spheres	Volume of all unique spheres/all spheres(Å^3)	Volume ratio Unique/all spheres
1br5_A crystal structure (resolution 2.5 Å)	100	7	576	2482	0.025
1il9_A crystal structure (resolution 3.1 Å)	100	8	507	2516	0.028
1rzo_C w/combined crystal template + SQWRL sidechains	93	8	495	2824	0.033
1rzo_C w/all SQWRL sidechains	93	9	676	2805	0.034
1abr_A w/all SQWRL sidechains	40	9	655	2513	0.023

Table 2: Comparison of docking score values at active sites and unique sites of target and near-neighbor proteins. All comparisons (average differences and R^2 correlations) are with respect to docking scores to our reference strain of ricin from the unique analysis.

PDB entry	% homology to ricin target, (resolution in Å)	Average difference in DOCK score at active site	R^2 of DOCK scores at active site	Average difference in DOCK score at unique site	R^2 of DOCK scores at unique site
Ricin structures					
1br5	100, (2.50)	1.6±2.7	0.78	0.14±3.1	0.72
1br6	100, (2.30)	0.084±1.3	0.95	0.19±1.3	0.95
1rzo	93, (2.63)	2.2±3.7	0.72	0.72±6.0	0.45
Near-neighbor structures					
1abr	40, (2.14)	4.7 ±5.7	0.08	28±8.5	0.0026
1mrj	38, (1.60)	2.4 ±3.4	0.66	25±9.4	0.018
1tfm	38, (2.80)	2.8 ±4.9	0.31	7.6±5.7	0.21