Beyond BLASTing: Tertiary and Quaternary Structure Analysis Helps Identify Major Vault Proteins

Toni K. Daly*, Andrew J. Sutherland-Smith, and David Penny

Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

*Corresponding author: E-mail: t.daly1@massey.ac.nz.

Accepted: December 24, 2012

Data deposition: Sequences used in this research — Accession numbers: Q62667, Q5EAJ7, Q4CUM2, Q4QJJ7, Q62774, P35240, Q62774, P35240, D2V5B9, D2W0Z9, D2UZF7, D2VSY6, D2VC38, D2VH38 are deposited in UniProtKB. The Rat crystal structure 2ZUO is deposited in the Protein Data Bank. All versions used are current as of January 12, 2013.

Abstract

We examine the advantages of going beyond sequence similarity and use both protein three-dimensional (3D) structure prediction and then quaternary structure (docking) of inferred 3D structures to help evaluate whether comparable sequences can fold into homologous structures with sufficient lateral associations for quaternary structure formation. Our test case is the major vault protein (MVP) that oligomerizes in multiple copies to form barrel-like vault particles and is relatively widespread among eukaryotes. We used the iterative threading assembly refinement server (I-TASSER) to predict whether putative MVP sequences identified by BLASTp and PSI Basic Local Alignment Search Tool are structurally similar to the experimentally determined rodent MVP tertiary structures. Then two identical predicted quaternary structures from I-TASSER are analyzed by RosettaDock to test whether a pair-wise association occurs, and hence whether the oligomeric vault complex is likely to form for a given MVP sequence. Positive controls for the method are the experimentally determined rat (*Rattus norvegicus*) vault X-ray crystal structure and the purple sea urchin (*Strongylocentrotus purpuratus*) MVP sequence that forms experimentally observed vaults. These and two kinetoplast MVP structural homologs were predicted with high confidence value, and RosettaDock predicted that these MVP sequences would dock laterally and therefore could form oligomeric vaults. As the negative control, I-TASSER did not predict an MVP-like structure from a randomized rat MVP sequence, even when constrained to the rat MVP crystal structure (PDB:2ZUO), thus further validating the method. The protocol identified six putative homologous MVP sequences in the heterobolosean *Naegleria gruberi* within the excavate kingdom. Two of these sequences are predicted to be structurally similar to rat MVP, despite being in excess of 300 residues shorter. The method can be used generally to help test predictions of homology via structural analysis.

Key words: homology modeling, BLAST, I-TASSER, RosettaDock, *Naegleria gruberi*.

Introduction

Our interest has included identifying features, proteins, and nontranslated RNAs that may date back at least to the Last Eukaryotic Common Ancestor (LECA). It is increasingly appearing that LECA already had quite a complex cellular and molecular structure (Kurland et al. 2006; Koonin 2010; Neumann et al. 2010). Of particular interest are the smaller untranslated RNAs found in ribonucleoproteins, including the spliceosome (Collins and Penny 2005) and eukaryotic ribosome (Steitz and Moore 2003), where RNA plays a critical catalytic role. Vaults are large oligomeric ribonucleoproteins conserved among a variety of species, many of which contain small untranslated RNAs (vault RNA [vtRNA]) (Stadler et al. 2009). Could the vault RNP date back to similarly early times? We need to be able to include structural information to test predictions made solely on linear (sequence) information. We first discuss the vaults, then the need for tertiary and quaternary protein structures to help the search for homology.

Vaults can be directly observed by electron microscopy or be detected by immunoblotting with anti-major vault protein (MVP) antibodies, in diverse species such as sea urchins (Hamill and Suprenant 1997), cellular slime mold (Vasu et al. 1993), electric ray (Herrmann et al. 1997), and mammals (Kedersha and Rome 1986). The rat vault RNP structure has been determined to 3.5 Å (Tanaka et al. 2009) defining both the MVP monomeric conformation and how 78 monomers assemble to form a complete vault (a half vault is shown in fig. 1*A*). The rat MVP monomer consists of four regions: multiple

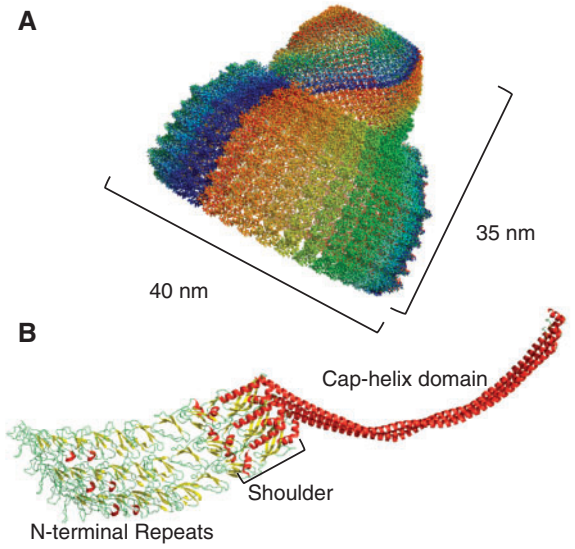

Fig. 1.—Vault ribonucleoprotein structure. (*A*) Rat MVP quaternary structure showing half a vault colored by monomer (PDB: 2ZUO, 2ZU4, and 2ZV5). A full vault will have at the lower left a copy of the upper half vault related by a 2-fold rotation axis. (*B*) Three rat MVP monomers colored by secondary structure (PDB 2ZUO stripped down to three monomers). This figure highlights the extensive lateral association required to dock into the vault quaternary structure.

N-terminal repeat domains, a shoulder domain, and the cap-helices (fig. 1*B*); additionally, there is a fourth domain, the cap-ring that is not sufficiently ordered to be observed in the crystal structure and so is not visible in figure 1. The rat monomer begins with nine repeat domains from the N-terminus—these repeats form a "stave-like" structure along the side of the vault barrel. The repeat domains are followed by the shoulder domain that then connects to a 42 turn α-helical domain known as the cap-helix. The cap-helix represents the top of the vault at a lower diameter than the N-terminal repeats (fig. 1*A* and *B*). Interactions between monomers of the long helical cap-helix are key for vault stabilization (Tanaka et al. 2009) and are essential for self-assembly (van Zon et al. 2002).

The equilibrium of monomer to oligomeric vault appears to strongly favor vault formation. For example, in rat liver cell lysate, ultracentrifugation of purified MVP shows 95% of the population as a high molecular weight form (Kedersha et al. 1991); and antibodies fractionate with intact vaults rather than with individual monomers in rat neural cells (Paspalas et al. 2009). Vaults are stable to a wide pH range (4–11), as well as in 1% Triton X-100 and 2 M urea (Kedersha et al. 1991). Extension at the N or C terminal does not prevent vault formation as fusion tagged MVP still assembles into vaults (Kickhoefer et al. 2009). Although vaults have other components, vtRNA, vault poly ADP-ribosylating protein (VPARP), and telomerase associated protein 1 (TEP1), these are not normally essential for vault formation (Stephen et al. 2001). TEP1 is also found in the telomerase complex; additionally, VPARP and vtRNA are found outside of vaults and so may have other functions as well. Although vault RNPs are linked to many processes (Berger et al. 2009; Vollmar et al. 2009; Lara et al. 2011; Liu et al. 2011) as yet they have no known intrinsic function.

The general issue of homology arises because proteins annotated as MVP via sequence homology, rather than by experimental determination, have been reported in the genome of many species including trypanosomes and paramecium. Considering that MVP sequences are apparently reasonably widespread, numerous and relatively conserved, it is surprising that convincing homologs (sequences or structures) appear to be missing from nematodes, flies, and fungi. A plant homolog recently reported in domestic barley (*Hordeum vulgare*) (Matsumoto et al. 2011) (UniProtKB: F2E078) has yet to be ascribed to the barley genome, thus could be the result of contamination—an example of contamination has been reported in mosses (Stevens et al. 2007). Thus, we require structural prediction information to help confirm (or not) the presence of vaults in a wider range of eukaryotes.

Traditionally, linear protein sequences have been used to determine homology, with subsequent annotation extrapolated to similar sequences based on a small subset of experimentally characterized proteins. Protein structure may sometimes be minimally affected by amino acid substitutions, and sequences with limited similarity may retain homologous folding patterns (Murzin et al. 1995; Orengo et al. 1997). In addition to sequence comparison, modeling studies have been used to identify members of protein superfamilies with low sequence homology (Holm and Sander 1997) and can also be used to predict function (Watson et al. 2005). Structural prediction studies are especially important for evaluation of the deepest sequence similarities because the Markov models we use for sequence evolution are expected to saturate, and lose information, at the most ancient divergences (Mossel and Steel 2004). Another way of testing structural and functional predictions is to synthesize the inferred ancestral sequences and measure their properties (Finnigan et al. 2012).

To extend one-dimensional sequence homology analysis, we have used a computational approach to help identify putative MVP sequences and to determine whether they are likely to form intact vaults. MVP represents an ideal case for the purposes of demonstrating the utility of three-dimensional (3D) studies as a means of enhancing the search for functional sequence homologs because the oligomeric vault structure is capable of independent self-assembly (Stephen et al. 2001). Furthermore, the monomeric MVP tertiary structure (and hence the sequence) is presumably under strong selective pressure to retain a conformation that forms not only the appropriate monomer structure but also the appropriate interface interactions with its neighbors for vault quaternary

structure/assembly (Qian et al. 2011). Here, we examine previously uncharacterized putative MVP sequences against these structural criteria, enabling us to predict with improved certainty whether the *mvp* gene, or relics of it, is likely to be present in a given species and whether intact vault particles are likely to form. Controls (both positive and negative) are essential to help determine the reliability of the inferred tertiary and quaternary models. It is essential to use tertiary and quaternary information to test homologies suggested in linear (one dimensional) information, and we have used many standard programs that are outlined later.

Electron microscopy of vault particles from a variety of species indicates that the intact vault structure is strikingly conserved. The rat MVP structure (PDB:2ZUO*b) was chosen as the standard by which we compare the folding of all other models because the whole oligomeric vault is resolved to 3.5 Å (Tanaka et al. 2009). Other structures in the Protein Data Bank (PDB) are fragments limited to the repeat sections of the MVP monomer only: mouse (Querol-Audi et al. 2009) and human (Kozlov et al. 2006), both virtually identical to the rat structure. The rat MVP structure is not necessarily an ideal template for the structure of distantly related MVP sequences, and the amoebozoa *Dictyostelium discoidium* forms a vault from a chimera of two structurally similar MVP paralogs (Vasu and Rome 1995). However, because it is the only full-length oligomeric vault structure, all comparisons have been made to the rat sequence and structure.

Materials and Methods

Basic Local Alignment Search Tool Searches

Initial Basic Local Alignment Search Tool (BLAST) searches were undertaken with the rat MVP accession Q62667 and Universal Protein Resource (UniProtKB) using the default BLOSUM62 matrix. All accession numbers refer to UniProt. Later searches used the less stringent BLOSUM45 to identify more remote sequences. "Expect values" (*E*) greater than 0.15 routinely produced BLAST matches that corresponded only to the cap-helix region of MVP (residues 647–802). Similarly, most PSI-BLASTs of the NCBI database identify false positives following the first iteration that align only with the coiled coil and no other MVP region. A PSI-BLAST search should not unduly weight the cap-helix region; however, it appears that there are a limited number of positional homologs involving the repeat areas and an abundance of proteins with the common coiled-coil motif. A search of conserved domains (National Center for Biotechnology Information) shows a very large overlap of conserved domains within the MVP cap-helix region—so the PSI-BLAST search was repeated without the inclusion of the cap-helix and using the kinetoplast sequence from *Leishmania major*. However, the cap-helix was restored following the first iteration as the 1,000 sequences retrieved (default is 500) are aligned, and a positional matrix is formed and used as the query for the second iteration. Similar searches were also undertaken using ancestral sequences reconstructed from 14 leishmania sequences and 14 trypanosome sequences, but no further sequences were found.

Iterative Threading Assembly Refinement Server

Iterative threading assembly refinement server (I-TASSER) inputs a query sequence and generates 3D structural models from multiple threading alignments using LOMETS (LOcal MEta Threading Server), a combination of eight threading programs (FUGUE, HHsearch, MUSTER, PROSPECT2, PPA, SP3, SAM-T02, and SPARKS) (Zhang 2008). The submitted sequence initially undergoes a PS-BLAST search to identify possible evolutionary relatives. I-TASSER then uses this BLAST result to generate a position-specific scoring matrix (PSSM or profile) using sequences with an E value lower than the threshold (0.005 is the default). The server uses this information to generate a PSI-BLAST using the PSSM as the query. It continues in this manner until no new sequences are added. Still within I-TASSER, the resultant profile is submitted to the PSIPRED server for secondary structure prediction, and both are then submitted to LOMETS. The final structure is presented by MODELLER (Sali and Blundell 1993) using a program that creates a probability density function using geometric criteria that satisfies spatial restraints within the query sequence in comparison to solved structures. It additionally has some ability to predict the shape of the loop structures, which, in the case of the vault, is useful for coverage of the sections missing from the experimentally determined structure (fig. 2*A*).

I-TASSER is benchmarked by Critical Assessment of Techniques for Protein Structure Prediction (CASP) (Moult et al. 1995), a biannual experiment in which servers are tested on their ability to identify correct folds from protein sequences whose structures have been previously determined but held back from publication by the PDB for the experiment. I-TASSER has scored highly since its inception competing as "Zhang Lab," winning best structure prediction and best function prediction in the most recent test in 2010 (Xu et al. 2011).

The most relevant score for the models predicted by I-TASSER is the *C* score with range −5 to +2. This is the confidence score for the estimated quality of the models calculated from the structural threading and refinement. A *C* score >-1.5 is considered to be a correct fold (Roy et al. 2010). The template modeling (TM) score quantifies structural similarity between two superimposed protein structures analogous to the traditional root mean-squared difference (RMSD). A TM score >0.5 indicates high confidence that the topology of two models, in this case predicted and native, is the same, and a TM score <0.17 indicates that the comparison is between random structures. The *C* score is correlated to the TM score (correlation coefficient 0.91) (Zhang 2008). TM weighs small

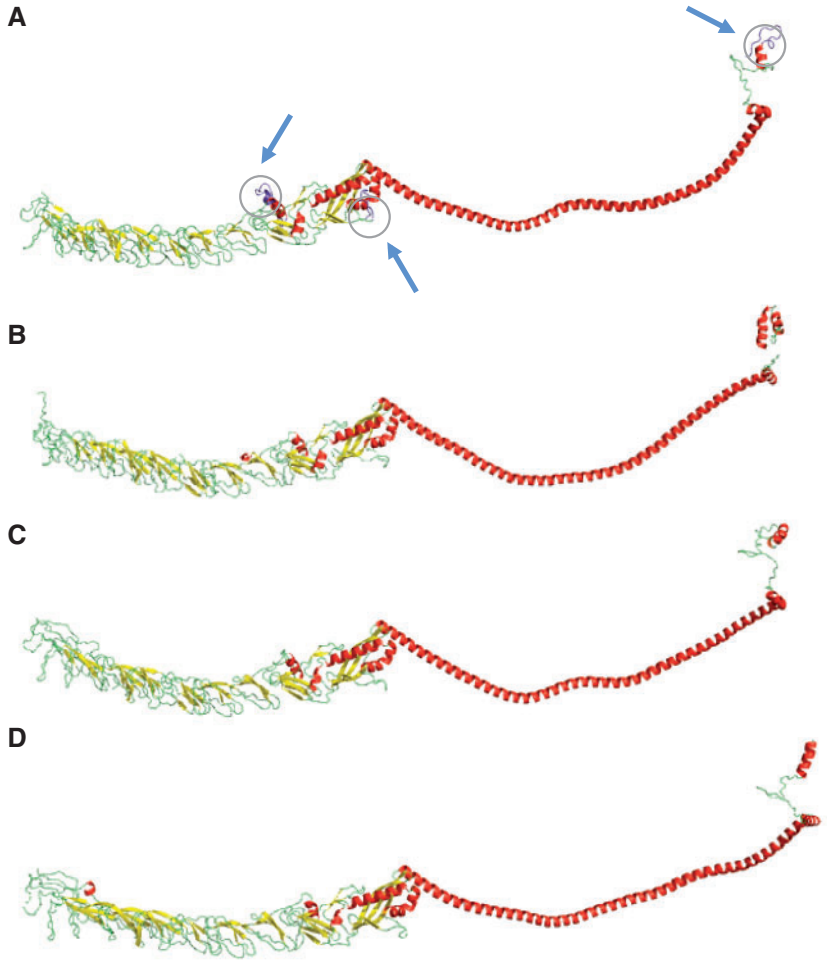

Fig. 2.—MVP monomer comparison. (*A*) I-TASSER-modeled structure for the full-length rat MVP sequence (Q62667). Residues not observed in the crystal structure (PDB:2ZUO*b) are circled (shown by arrows). (*B*) I-TASSER-modeled structure for the sea urchin MVP monomer (Q5EAJ7). (*C*) I-TASSER-modeled structure for the kinetoplasts *Trypanosome cruzi* (Q4CUM2) and (*D*) *Leishmania major* (Q4QJJ7) MVPs.

distance differences greater than large ones and has a length-dependent normalization scale. In contrast, RMSD weighs the pair-wise differences between residues equally meaning that a local difference can have a large impact on the RMSD score, particularly if the protein is large. Because MVP is approximately 850 residues, the RMSD is likely to be of less value. The final control for model quality before submission to RosettaDock was visual comparison to the rat structure, because the "I-TASSER best model" was not necessarily the one that looked most closely like a vault monomer.

Although we considered each output from I-TASSER on a case-by-case basis, some general criteria were applied. For example, to choose a model visually rather than because it is the result with the highest *C* score, the *C* scores of the models concerned must be similar. If the *C* scores are similar, as may occur for targets described by LOMETS as "hard," the first model presented by I-TASSER is not necessarily the best (Roy et al. 2011). Additionally, the *C* score information lists the number of decoys and cluster density for each output. If these are also similar for the models being compared, then the model is chosen that is visually closest to the known structure.

If the target is described as "easy," then the first model generally has a significantly higher *C* score than the rest. LOMETS produced a variety of structures from the *Naegleria gruberi* MVP sequences found via the PSI-BLASTs described as "medium" targets. Visually they were all different, none looked like MVP, and although LOMETS alone does not give a score for confidence, the probability that the models showed the correct folds was described as "medium." They were then submitted to I-TASSER using the rat crystal structure (2ZUO*b) as a constraint. When a constraint is used, it can be applied with or without a specified alignment. If an alignment is not specified, then the MUSTER (MUlti-Source ThreadER) algorithm is used (Wu and Zhang 2008). The initial full-length rat MVP sequence shown in figure 2*A* could have been used

as a constraint that would have resulted in greater uniformity, particularly with respect to the C terminal and amorphous loop on repeat eight. However, by using 2ZUO*b, it means that I-TASSER has repeatedly modeled the missing residues ab initio. In fact, the amorphous loop makes a shelf on the inside of the vault that is consistently modeled by RosettaDock. Because the most C-terminal region was not visible in the crystal structure, yet is very highly conserved, future evaluation of the models should highlight any consensus folds in this area. We additionally confirmed the predicted folds using Phyre^2.

Phyre^2

Phyre^2 is an upgrade to the original Protein Homology/analogY Recognition Engine (Phyre) (Kelley and Sternberg 2009). Phyre takes a sequence, builds a profile using PSI-BLASTs, and compares it to templates deposited in the Structural Classification of Proteins database and PDB. Phyre uses three secondary structure prediction programs: PSIPRED, SSPro, and JNet. Each program gives a confidence value for each of three structures: alpha helix, beta sheet, and coil. The confidence values are averaged, and a final, consensus prediction is displayed for each individual prediction. This is computationally less expensive than the multiple alignments used by I-TASSER generating much quicker results and has the advantage that multiple, or even "batch," submissions can be made. Additionally, 20 results can be displayed in full and many more suggested, which means that individual folds can be identified. Phyre and Phyre^2 have been similarly successful in the CASP experiments.

Flexible Structure Alignment by Chaining Aligned Fragment Pairs Allowing Twists

Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists (FATCAT) (Ye and Godzik 2003) gives a measure of similarity of one structure to another. Structural models predicted by I-TASSER from query sequences were compared with the rat MVP monomer. FATCAT breaks the proteins to be aligned into fragments eight residues long (aligned fragment pairs [AFPs]). These AFPs can be matched, and a twist, gap, or extension can be introduced to match the next AFP if it results in a substantially better superposition. Extensions, gaps, and twists are all scored using a dynamic programming algorithm, so that long AFPs are rewarded and large RMSDs are penalized. This gives the lowest possible chaining score at each juncture. The total chaining score is then combined with the probability of obtaining a greater score, the RMSD of the final superposition, the number of equivalent positions, and the number of twists (with a maximum of five), to give a measure of the structure's significance. This is displayed both as a *P* value and as a raw score. When comparing MVP models, the *P* value is most often reported as "zero," so the raw score gives a sense of "more" or "less" similar to the rat structure—a high raw score indicates greater similarity to the rat crystal structure that it is being compared with (data not shown).

In this instance, FATCAT was used to space the MVP monomer models for RosettaDock analysis by aligning the query structures with 2ZUO*b and with 2ZUO*d (i.e., one monomer width apart) of the rat crystal structure. In some cases, FATCAT will introduce chain breaks to undo twists in the aligned models making them unsuitable for docking analysis; FATCAT can be forced to run a "rigid" alignment that will prevent breaks, and this is a simple and almost instantaneous way of suitably spacing the monomers. Another approach used was to manually position the molecules a monomer width apart in PyMOL (The PyMOL Molecular Graphics System, Version 1, DeLano scientific LLC, 2008), although the advantage of using FATCAT was that the RMSD could be predicted and thus help identify possible docked models where scores were similar across the majority of the models.

RosettaDock

For vault formation, the MVP monomers dock laterally along the length of both sides to make the barrel shape. RosettaDock is a server that uses a low-resolution Monte Carlo search and backbone optimization algorithm to position the submitted chain pair, followed by a refinement to relax the backbone and accommodate the side chains (Gray et al. 2003).

RosettaDock has very specific requirements; two monomers, side by side, are submitted to see whether they will dock laterally. If the pair of monomers input for docking are initially placed too far apart, then the first local docking search performed may fail to locate them. However, if they are placed too close together (<5 Å), the file is rejected. Additionally, the RosettaDock file cannot total more than 600 residues for submission to the online server as such calculations are computationally too expensive. RosettaDock can be downloaded as a package and thereby the number of residues can be increased. For the online server, the MVP monomers were docked in three sections. The cap-ring domain (C terminal ~60 residues) has not been submitted to RosettaDock because, although it is highly conserved, there is no suitable experimentally determined control structure. As a final complication, in some instances, RosettaDock docks MVP monomers with a large energy score skewed by internal residues that are not involved with the oligomerization interface.

To benchmark RosettaDock, other servers have been tried; ClusPro (Kozakov et al. 2006) is unable to take such large regions of MVP due to a 24-h job limit. GrammX (Tovchigrechko and Vakser 2006) is considerably quicker than Rosetta, but in some instances, it docked the N terminal

of the vault proteins in an antiparallel orientation, which is not consistent with the oligomeric vault crystal structure.

Results

Positive Control Study for Method Optimization: Tertiary Structure

The first control used the rat MVP sequence (Q62667) (Kickhoefer and Rome 1994) to model the MVP monomer structure via the I-TASSER server (Roy et al. 2010), initially unconstrained, then constrained by the rat crystal structure (PDB:2ZUO*b) (Tanaka et al. 2009). This confirmed that I-TASSER identified correctly the crystal structure from the full-length rat sequence. The rat MVP crystal structure shows only 812 residues of the total 861 amino acid sequence. Three regions not observed in the crystal structure are residues 429–448 (a presumed disordered loop on repeat 8), 608–620 (part of the shoulder domain), and amino acids 846–861 (the very C terminus, beyond that described as the cap-ring domain). Nevertheless, the I-TASSER prediction for these regions is important because I-TASSER will be modeling full-length homologous MVP sequences of unknown structure (fig. 2*A*). FATCAT structural alignment showed generally that the predicted model is very close to the experimental crystal structure regardless of whether the I-TASSER input sequence was constrained to the known rat structure.

As an additional control, the MVP sequence from the purple sea urchin (an echinoderm), *Strongylocentrotus purpuratus* (Q5EAJ7) was analyzed. This urchin MVP has 64% sequence identity with the rat, and intact vaults have been seen via cryo-electron microscopy (Stewart et al. 2005), but the urchin MVP does not have a crystal structure determined. The urchin MVP sequence was submitted to I-TASSER without 2ZUO*b constraint, and the resulting fold (fig. 2*B*) is very similar to that of the rat (fig. 2*A*). MVP sequences from the kinetoplasts *Trypanosome cruzi* (Q4CUM2) (fig. 2*C*) and *L. major* (Q4QJJ7) (fig. 2*D*) were also analyzed (unconstrained) to model the structure that could be anticipated for excavate MVPs. Results are reported in table 1.

All sequences were also submitted to I-TASSER using 2ZUO*b (from the rat crystal structure) as a constraint to determine the influence a structural constraint has on the modeling. The use of this constraint has no discernable effect on

Table 1
I-TASSER and RosettaDock Results for Positive and Negative Controls

UniProtKB Accession Number	Organism	Length	% Identical Sites versus Q62667	I-TASSER *C* Score	I-TASSER TM Score	RosettaDock Score for Cap-Helix	RosettaDock Score for Shoulder and Cap-Helix
Positive controls, unconstrained							
2ZUO*b	*Rattus norvegicus*	812				−261	−435
Q62667	*R. norvegicus*	861	100	0.42	0.77 ± 0.10	−280	−254
Q5EAJ7	*Strongylocentrotus purpuratus*	857	64	1.12	0.87 ± 0.07	−291	−503
Q4CUM2	*Trypanosome cruzi*	838	48	1.11	0.87 ± 0.07	−304	−498
Q4QJJ7	*Leishmania major*	833	48	1.91	0.99 ± 0.04	−302	−504
Positive controls, constrained by 2ZUO*b							
Q62667	*R. norvegicus*	861	100	1.02	0.85 ± 0.08	−292	−508
Q5EAJ7	*S. purpuratus*	857	64	1.07	0.86 ± 0.07	−255	−492
Q4CUM2	*T. cruzi*	838	48	1.18	0.88 ± 0.07	−247	None docked
Q4QJJ7	*L. major*	833	48	1.33	0.90 ± 0.06	−266	None docked
Negative controls, unconstrained							
Randomized rat MVP	*R. norvegicus*	861	16	−1.76	0.50 ± 0.15	No cap-helix	—
Q62774	*R. norvegicus* (myosin 1A)	842	16	0.96	0.84 ± 0.08	−258	No shoulder
P35240	*Homo sapiens* (merlin)	595	17	−0.76	0.62 ± 0.14	No cap-helix	—
Negative controls, constrained by 2ZUO*b							
Randomized rat MVP	*R. norvegicus*	861	16	−2.93	0.38 ± 0.13	No cap-helix	—
Q62774	*R. norvegicus* (myosin 1A)	842	16	0.62	0.80 ± 0.09	−191	No shoulder
P35240	*H. sapiens* (merlin)	595	17	−1.33	0.55 ± 0.15	Helix does not dock	

Note.—The I-TASSER confidence (*C*) score (>−1.5 is considered a correct fold, range −5 to +2, higher is better). The RosettaDock energy score is lower for the shoulder and cap-helix combined, indicating that the shoulder improves docking. It should be noted that the lateral docking capacity of the cap-helix in the rat MVP was reduced in comparison to the other positive control sequences (fig. 3). This was improved by using the 2ZUO*b constraint for the I-TASSER rat MVP prediction. In general, using the constraint during I-TASSER modeling reduced the likelihood that RosettaDock would successfully dock the modeled monomers. The other positive control MVP sequences were also submitted to I-TASSER constrained by the rat crystal structure 2ZUO*b. With the exception of rat and *L. major*, this made very little difference to the I-TASSER score, but it did reduce the possibility of finding docked monomers in the excavates. In the case of the rat, both I-TASSER and RosettaDock scores are considerably improved by using the 2ZUO*b constraint. The score for *L. major* is reduced by the 2ZUO*b constraint but still well above the threshold of confidence that the model is correct. (*B*) Comparison between negative control I-TASSER models with and without the 2ZUO*b constraint shows that the constraint does not make I-TASSER any more likely to find that the structure matches the rat crystal structure template but does lower the confidence that LOMETS has in the resulted structure. The high *C* score for the rat myosin (shaded) reflects the myosin V (PDB 2DFS) database structure identified by I-TASSER as most similar. The low score for the randomized rat MVP sequence reflects little similarity to any of the structures in the PDB.

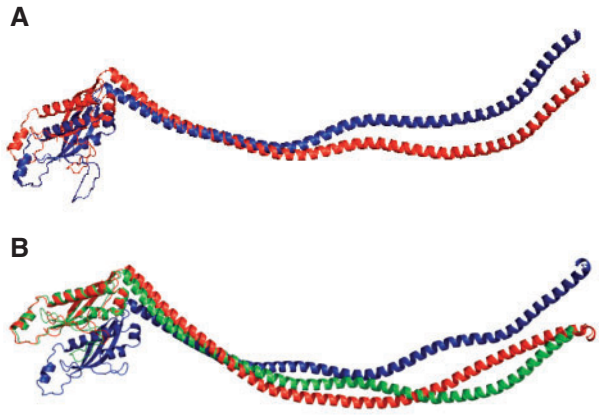

Fig. 3.—Structural effect of the 2ZUO*b constraint. (*A*) Structural comparison of the shoulder and cap-helix region of two rat MVP models either constrained by 2ZUO*b (red) or unconstrained (blue). The kink in the unconstrained cap-helix modeled by I-TASSER results in poor docking in RosettaDock. The rat MVP sequence constrained by 2ZUO*b (red) entirely aligns with 2ZUO*b (obscured), and this model docks readily in RosettaDock. (*B*) Urchin MVP shoulder and cap-helix region structural comparison between models either constrained by 2ZUO*b (red) or unconstrained (blue) relative to 2ZUO*b (green). In this case, the unconstrained urchin MVP model docks more readily than the constrained model.

the rat, sea urchin, and kinetoplasts sequences in terms of the repeat and shoulder domains. However, the cap-helix structures were altered by the constraint, which had a subsequent effect on the docking performance of the structures (fig. 3 and table 1). All sequences were additionally submitted to the Phyre^2 protein fold recognition server. Phyre^2 confirmed the I-TASSER results with 100% confidence (data not shown).

Positive Control Study for Method Optimization: Quaternary Structure

As a further control to determining whether the putative MVP sequences fold in a similar manner to the characterized rat monomer structure, we need to ascertain whether sequences with high structural homology to the MVP monomer are likely to dock with each other and form a vault. As a control, we analyzed rat MVP monomer structures with RosettaDock, with either MVP monomers taken directly from the crystal structure or the full-length rat monomeric MVP structure predicted by I-TASSER. RosettaDock predicts an oligomeric vault structure similar to that of the crystal structure, which can form with good low energy scores (table 1). The MVP C-terminal long α-helix has been shown to be essential for self-assembly of monomers into oligomeric vaults (van Zon et al. 2002). Therefore, the cap-helix regions (amino acids 647–802) of two separated rat MVP monomers from the crystal structure (PDB:2ZUO*b and 2ZUO*d) were submitted to RosettaDock to test how well it would reassemble the lateral associations of docked MVP pairs required for vault assembly.

In most animal species, vaults are homo-oligomeric complexes constructed from identical MVP monomers, so the interactions between monomers are all the same. This means that the docking of one monomer pair can be used to infer vault formation if the appropriate lateral association forms. RosettaDock considers 1,000 structures and searches for the lowest energy conformations of which 10 are output. Each docking solution has an overall energy score (RosettaDock energy score, *y* axis) that is plotted against the RMSD (*x* axis) from the starting positions (Å) of the monomers. Score graphs showing a characteristic "funnel" suggest that the 1,000 pairs are clustered in conformation, giving a higher confidence in the lowest energy docked pairs resulted (Lyskov and Gray 2008). A score graph showing the energy scores versus RMSD for residues 647–802 from monomers of the rat crystal structure is shown in figure 4*A*, together with a cartoon (fig. 4*B*) and a surface rendered (fig. 4*C*) representation of the lowest energy docked pair of MVP monomers. Docked monomer surface and MVP ribbon representations were rendered with PyMOL.

A lower energy score can be found when the shoulder region (502–646) is included, indicating that the shoulder area probably contributes to the proper alignment and docking of the monomers (supplementary material S1, Supplementary Material online) consistent with the rat MVP crystal structure. Using the rat shoulder alone indicates a high probability that the shoulders will interact (fig. 5). Oligomerization of a domain homologous to the MVP shoulder has been experimentally demonstrated (Kuwahara et al. 2009).

Using the MVP domains separately, we show that the monomers are likely to dock along their entire length, even when missing the stabilizing effect of the coiled coil (supplementary material S1, Supplementary Material online) again consistent with the interdomain contacts identified from the MVP crystal structure. In each case, the energy score is low and negative, and the RMSD shows that the distance from the starting structure is well clustered. Because the monomers submitted to RosettaDock have been spaced by FATCAT one monomer width apart, the starting distance between the molecules is approximately 15–20 Å and the resulting RMSD for successful docking can be predicted. Thus, we test both that the modeled monomer MVP 3D tertiary structures are consistent with the rat MVP monomer structure and that those modeled monomers are likely to assemble into vaults.

Within the vault, MVP monomers contact their adjacent monomers laterally, but vaults are also able to open in a petal-like fashion from their equator (Kedersha et al. 1991; Yang et al. 2010), potentially complicating the docking analysis. Indeed, less than a third of the lateral noncovalent interactions between MVP monomers in the vault occur between the N-terminus and the shoulder domain (residues 1–519) with oligomerization dominated by interactions between the C-terminal cap-helix regions (van Zon et al. 2002;

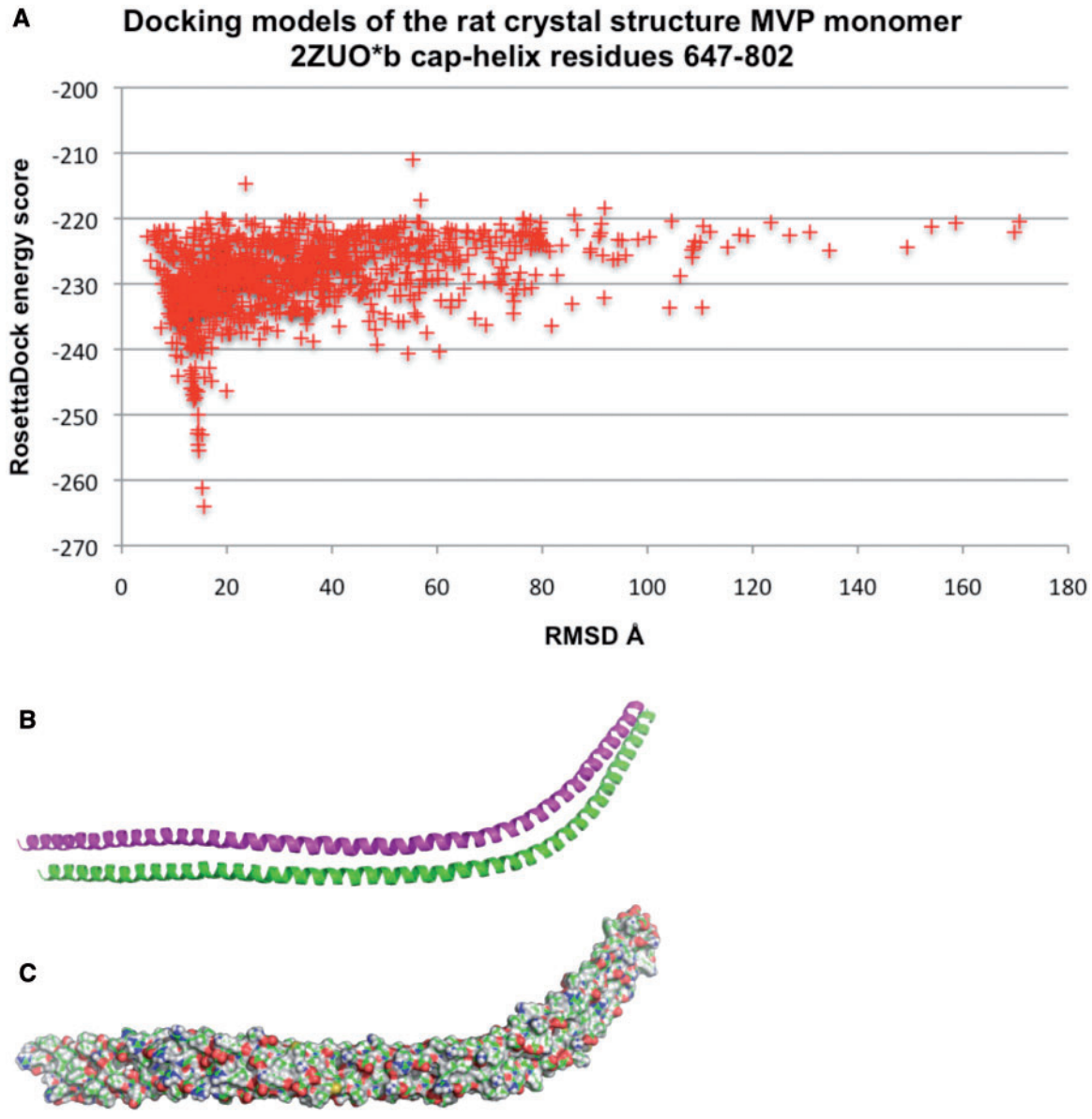

Fig. 4.—RosettaDock results from the crystal structure cap-helix. (*A*) Score graph depicting RosettaDock energy score versus RMSD (Å) of the docked monomers compared with their starting positions. The funnel shape of the score graph indicates a high confidence in the structure of the models with lowest energy score. (*B*) Cartoon of the lowest energy model (energy score −264) shaded by monomer. (*C*) Surface rendering of the lowest energy model.

Tanaka et al. 2009). This is demonstrated for the rat crystal structure MVP monomer by less favorable RosettaDock energy scores for the docking of the N-terminal sections of the monomer compared with the C-terminal shoulder and cap-helix consistent. In the case of 2ZUO*b, all 10 top models were docked along the length of the monomer. The RosettaDock output files list the pair energies across the interface; one of the 2ZUO*b cap-helix models showed residues paired as described for the crystal structure (Tanaka et al. 2009). However, the other RosettaDock output models, even those that included the shoulder—which could be expected to align the helix in position, showed various pairings. This indicates either some redundancy in the docking arrangements between the monomers in the shoulder and cap-helix or a lack of fine resolution in the RosettaDock prediction—given that the residues that interact across the oligomerization interface (identified in the crystal structure) are well conserved (see MVP sequence alignment marked with known interactions, supplementary material S2, Supplementary Material online). The remaining MVP positive control sequences were analyzed in the same way (table 1 and supplementary material S1, Supplementary Material online), predicting the formation of vault particles. For the other positive controls, including full-length I-TASSER-modeled rat MVP, the RosettaDock

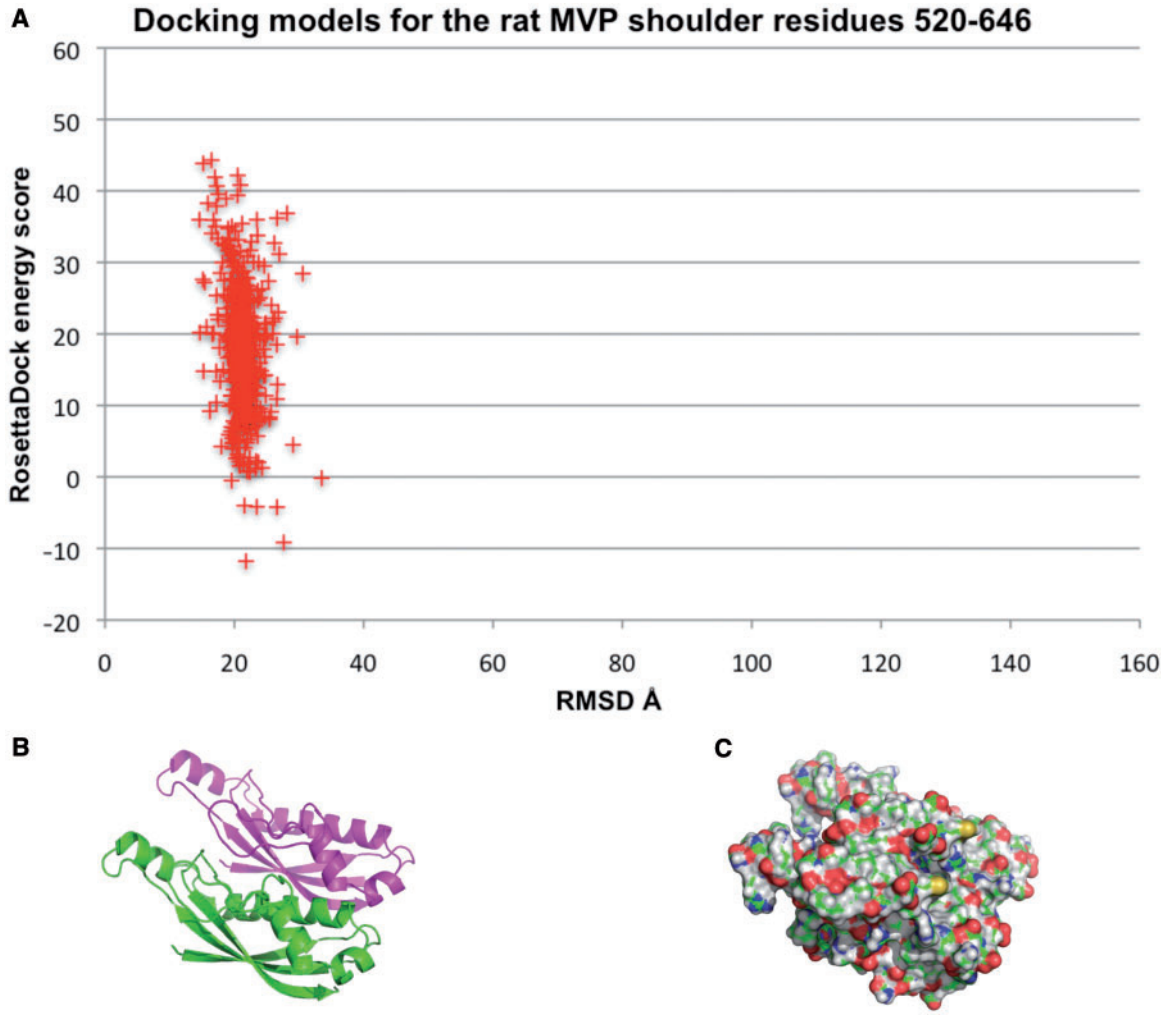

Fig. 5.—RosettaDock results from the rat MVP shoulder region. (*A*) Score graph representing the RosettaDock energy scores versus RMSD (Å) for the 1,000 models generated by RosettaDock for the shoulder region of MVP (residues 520–646). The energy score for the shoulder region docking is higher than for the cap-helix (table 1). (*B*) Cartoon of the shoulder domain from the lowest energy model of the two docked monomers (energy score −12) shaded by chain. (*C*) Surface rendering of the lowest energy docked monomers.

energy score for the repeat sections was significantly lower, that is, more favorable than the shoulder/cap-helix regions (table 1). Thus, the results are consistent with I-TASSER and RosettaDock being able to detect genuine vaults.

Negative Controls

As a negative control, the full-length rat MVP sequence was randomized in three fragments: repeat domains, shoulder domain, and cap-helix. Randomization was confined within each fragment to determine whether the cap-helix was having an undue influence regarding the I-TASSER modeling, because this region strongly influenced the BLAST results (mentioned earlier). Two remote sequences found in BLAST searches were used as additional negative controls: rat myosin 1A (a similar sized protein to MVP) (Q62774) that does not have an experimentally determined structure and human merlin, (P35240) a neurofibromatosis-2 tumor suppressor that has the structure of its FERM domain determined (PDB 3U8Z). All sequences were subject to the same protocol and submitted to I-TASSER with and without constraints to the rat crystal structure 2ZUO*b.

As expected, the randomized rat MVP sequence could not be modeled on any existing structural template with confidence. The top scoring models, based on human importin β (PDB 1QGR), were of low confidence (table 1; *C* score −1.76 and constraint by 2ZUO*b reduced this to −2.93) and so not considered a "correct fold" by I-TASSER. The rat myosin 1A sequence was identified as most structurally similar to the inhibited state of myosin V (PDB 2DFS) with reasonable confidence regardless of the 2ZUO*b constraint (*C* score 0.96, and 0.62 with constraint) (table 1 and fig. 6*B*). Additionally, Phyre[2]

could not report a model for the randomized rat sequence and also identified myosin V as the most similar template for the myosin 1A sequence.

However, using the 2ZUO*b constraint did influence the structural prediction for the merlin protein sequence (fig. 6*C* unconstrained, and fig. 6*D* constrained, by 2ZUO*b). Although there is a crystal structure for the FERM domain, I-TASSER predicted the unconstrained sequence to be more similar to the merlin homolog in the armyworm caterpillar (PDB 2ILKA) presumably because this is full length rather than the 300 residues of the FERM domain. The shoulder domain in the 2ZUO*b constrained prediction does look very similar to the MVP shoulder, which was identified as similar to the stomatin core of *Pyrococcus horikoshii* (Tanaka et al. 2009) (see fig. 6 insert). Phyre^2 identified the merlin sequence specifically as moesin (the fourth part of the FERM domain) from the armyworm (PDB 2ILJA) as their first rated sequence, although the human merlin FERM domain was identified with 100% confidence and 100% coverage but presumably not given the top rating because the sequence was significantly longer than the PDB structure.

Because I-TASSER did not predict that a coil, similar in any way to the cap-helix, would form with the randomized rat MVP sequence, RosettaDock modeling was not carried out. However, the rat myosin 1A was predicted to form a coil structure similar to MVP, so this modeled structure was aligned via FATCAT to monomer positions b and d of the vault complex and submitted to RosettaDock. In this case,

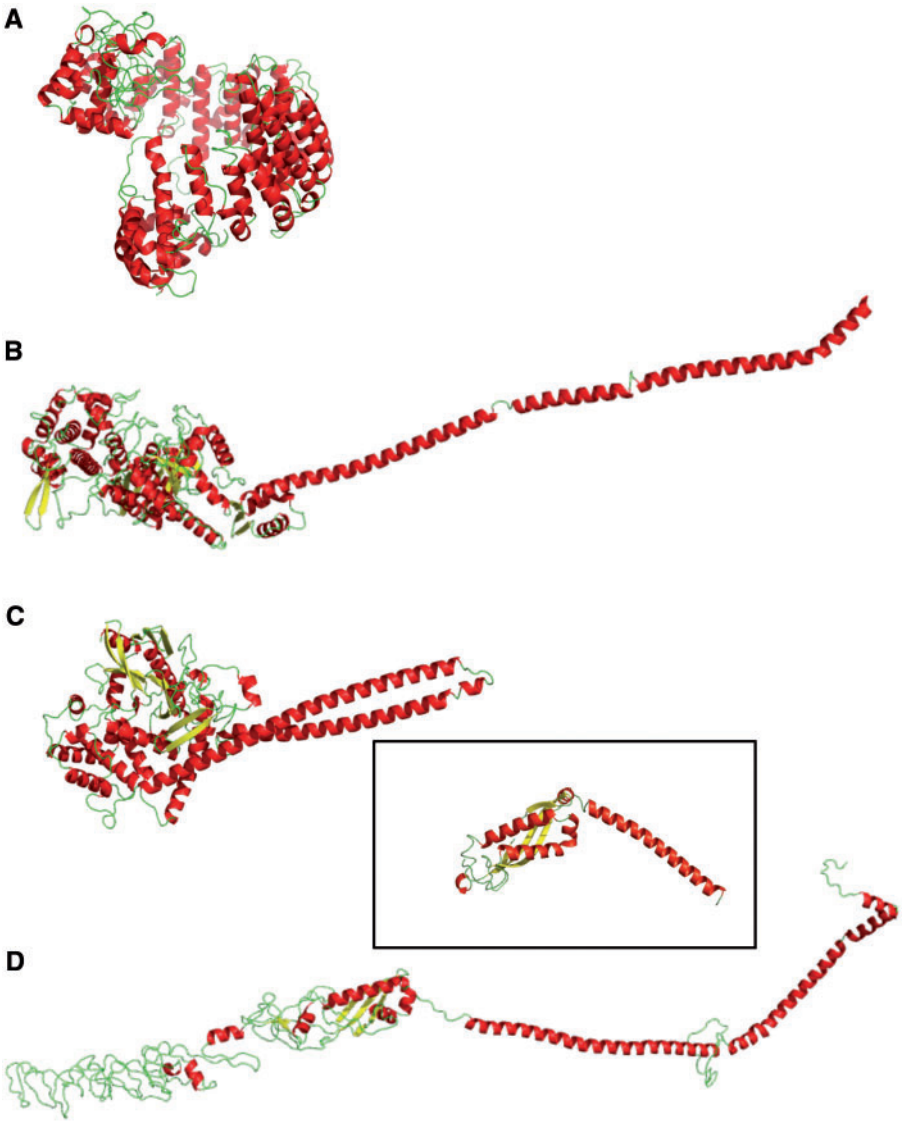

Fig. 6.—I-TASSER modeling results for the negative control sequences. (*A*) Randomized rat MVP. (*B*) Rat myosin 1A. (*C*) Human merlin unconstrained. (*D*) Human merlin constrained by 2ZUO*b. Insert is the stomatin core from *Pyrococcus horikoshii*.

the lowest energy model did dock along the length of the coiled coil, residues 675–815 (supplementary material S3, Supplementary Material online). The myosin motor domain was submitted to RosettaDock, and this also docked along its length, though with much higher (and positive!) energy scores (lowest energy score +1,020). The putative cap-helix for the 2ZUO*b constrained human merlin model only partially docked, due to an interruption in the coiled structure (residues 476–513). The shoulder area and truncated cap-helix (residues 315–475) were resubmitted and docked with an energy score of −113. Residues 1–407 representing a combination of the shoulder and the relatively unstructured sequence (in comparison with the repeated β sheets of MVP) were also predicted to dock laterally along its entire length though with a high energy score of +573 (see supplementary material S4, Supplementary Material online). This demonstrates that not all proteins with some homology with MVP (as they were retrieved using BLAST) will be predicted to fold similar to MVP even using the known rat crystal structure as a constraint. In the case of the merlin protein, where the rat constraint did influence the structures output by I-TASSER, it was then very difficult to dock identical monomers in RosettaDock. Thus, it is important that a suite of approaches is used to test structural homology.

Investigation of MVP Sequences from *N. gruberi*

Next we used the protocol to find MVP sequences in other genera. Initial BLASTp searches resulted in hundreds of putative MVP sequences, which were reduced to a data set of those with *E* value reported as "zero" and of a similar length (~850 residues) to the complete rat MVP sequence. No sequences matching these criteria were found from the ecdysozoa, or from fungi, but some were from kinetoplasts (excavates), some oomycetes (stramenopiles), and paramecium (an alveolate). With the criteria relaxed to include sequences with an *E* value up to 10, and any length, then the most remote (compared with rat) excavate sequence that has any kind of MVP annotation was found in *N. gruberi,* an excavate of the clade Heterolobosea, thought to be a very anciently diverged free-living protist. *Naegleria gruberi* has two putative MVP-like protein sequences with an initial PfamA (Finn et al. 2010) annotation of an "MVP shoulder domain" (UniProtKB:D2V5B9, which may not be complete, and D2W0Z9, which is described as "complete"). These two sequences are considerably shorter, 559 and 530 residues, respectively, and contain 17% (148/861) and 19% (166/861) identical sites compared with rat MVP. The size difference is mainly in the body of the vault with *N. gruberi* having fewer repeats domains, suggesting that either repeats have been gained in metazoa since their ancestors diverged from Heterolobosea or that *N. gruberi* has lost a region of the gene within the repeat section compared with the longer characterized MVP sequences. The sequence similarity between these two *N. gruberi* proteins is 35%, indicating that they have been evolving independently for a long time. If the rat MVP repeat region sequence is truncated in an equivalent manner, the percentage of identical sites rises to 25% in both cases (148/588).

The free living *N. gruberi* is often considered to be a representative genome present at a very early stage of eukaryote evolution (Fritz-Laylin et al. 2010). It is predicted to have 15,727 protein coding genes, 3,784 of these are found in at least three other eukaryotic supergroups and a further 349 are found in at least one other supergroup. In contrast, parasitic protists have a reduced genome, relative to their ancestors, owing to their lifestyle. I-TASSER modeled the *N. gruberi* putative MVP sequences into MVP folds with high TM and *C* scores both unconstrained and constrained by the 2ZUO*b template (table 2 and fig. 7). In both instances, the models

Table 2

I-TASSER and RosettaDock Results for the *Naegleria gruberi* Sequences

UniProtKB Accession Number	Length	% Identical Sites versus Q62667	I-TASSER *C* Score	I-TASSER TM Score	Rosettadock Score For Cap-Helix	RosettaDock Score for Shoulder and Cap-Helix
Sequences submitted to I-TASSER without constraint						
D2V5B9	559	17	−0.74	0.62 ± 0.14	−227	−441
D2W0Z9	530	19	0.07	0.70 ± 0.12	−226	−113
Sequences submitted to I-TASSER constrained by 2ZUO*b rat crystal structure						
D2V5B9[a]	559	17	0.98	0.85 ± 0.08	−209	−438
D2W0Z9	530	19	−0.26	0.68 ± 0.12	−287	−113
D2UZF7	845	13	−1.29	0.55 ± 0.15	No cap-helix	—
D2VSY6	833	13	−2.03	0.56 ± 0.15	None dock	None dock
D2VC38	694	16	−0.24	0.72 ± 0.11	−197	None dock
D2VH38	418	13	−3.15	0.36 ± 0.12	−165	None dock

NOTE.—For D2V5B9[a] constrained by 2ZUO*b, the lowest 10 energy score models did not dock. Docked models were identified from the expected RMSD and were 47th and 45th lowest energy, respectively. In both cases, the energy scores were all very similar, and there was no compelling consensus model (see supplementary material S5, Supplementary Material online). The highlighted gray cells are scores for I-TASSER predictions that do not resemble the MVP fold being structurally similar to human importin β.

(fig. 7*A* and *B*) clearly resembled the MVP structures from figure 2.

Because trypanosomes and leishmania have multiple copies of MVP homolog, it was hypothesized that *N. gruberi* may also have sequences not found by BLASTp, so a PSI-BLAST was conducted (Altschul et al. 1997) (Schäffer et al. 2001) using the first 625 residues of the *L. major* control sequence Q4QJJ7 as the query (see Materials and Methods). Four more *N. gruberi* sequences were retrieved with limited similarity (maximum 16%) to either rat or *L. major* MVP (table 2). All these were first submitted to LOMETS rather than I-TASSER in the interests of speed, but none of the resulting models predicted a structure that resembled MVP. As a further test, the sequences were submitted to I-TASSER constrained by 2ZUO*b (fig. 7*C–F*).

Of these additional sequences, only D2VC38 (694 residues; fig. 7*E*) is modeled by I-TASSER to resemble MVP with a *C* score indicating confidence in the model. Although this is the second-"best" model from I-TASSER, the *C* score is equivalent to the first model, and the cluster density is similar for both models with a similar number of decoys, meaning that the distinction between the two structurally dissimilar models is not certain (detailed in Materials and Methods). Although the model resembles MVP, there are clearly β sheets absent from

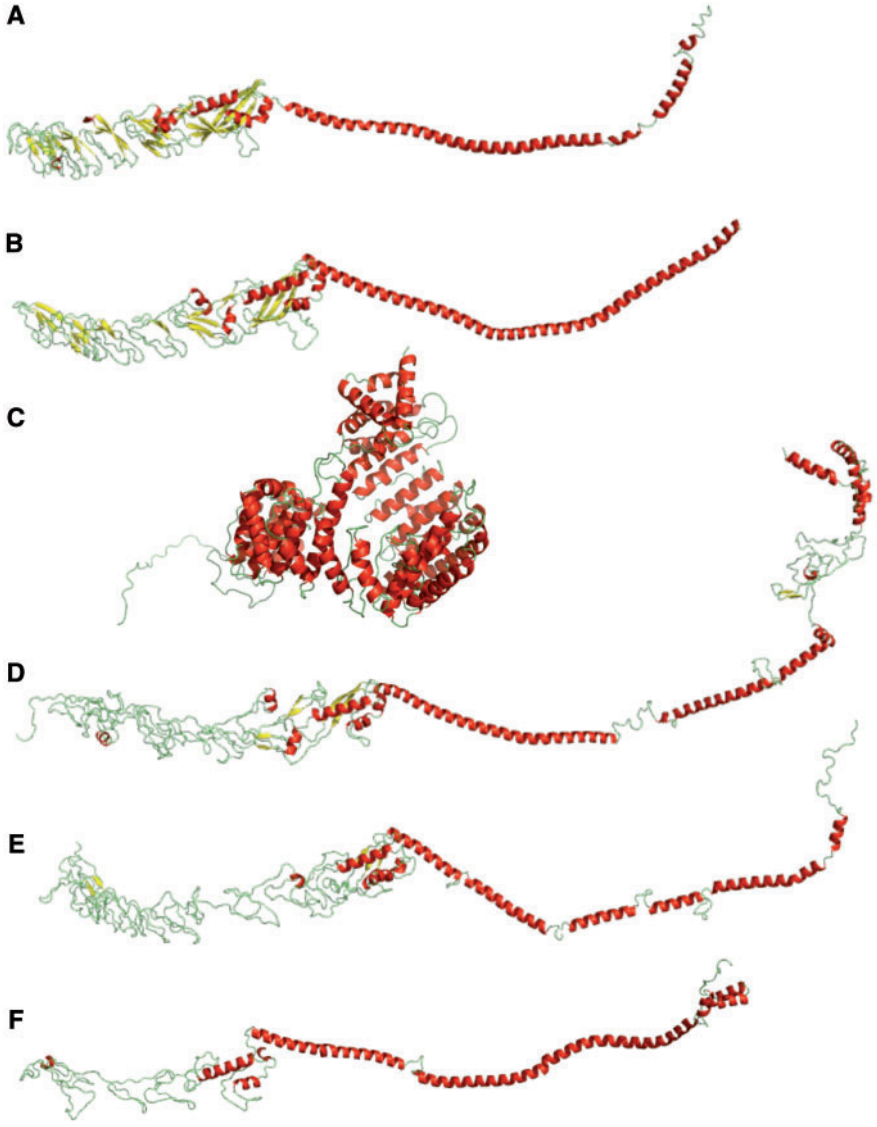

Fig. 7.—*Naegleria gruberi* MVP I-TASSER structural modeling. (*A*) D2V5B9, 559 residues. (*B*) D2W0Z9, 530 residues both identified from a BLASTp search of the UniProtKB database and submitted to I-TASSER without constraint. (*C–F*) Models derived from sequences retrieved via a PSI-BLAST of the National Center for Biotechnology Information (NCBI) database and submitted to I-TASSER constrained by the rat crystal structure 2ZUO*b. (*C*) DZUF7, 845 residues. (*D*) D2VSY6, 833 residues. (*E*) D2VC38, 694 residues. (*F*) D2VH38, 418 residues. UniProt accession numbers are provided for consistency. See also table 2.

the repeat domains. Sequence D2VSY6 (833 residues; fig. 7*D*) was also modeled resembling MVP. However, in this instance, the *C* scores of all the models are considerably lower and with a greater difference between the first and second models. The cluster density between these two models is also lower indicating that this prediction is probably no more likely than a random prediction. It could be that the extra C-terminal residues have contributed to the poor *C* score, even though experimentally a vault can still form with additional C-terminal residues. Interestingly, Phyre[2] identified the shortest sequence (D2VH38) as the bacterial transmembrane protein colicin Ia. A sequence identified as a "colicin uptake transmembrane protein" found in cyanobacteria *Lyngbya majuscula* (F4Y3B4) has 54% sequence homology with rat MVP and is predicted to fold identically to MVP by I-TASSER and Phyre[2]. F4Y3B4 is annotated by family and domain databases Pfam, InterPro, and PROSITE as MVP (see Discussion).

Quaternary Structure Prediction for *N. gruberi* Sequences

The *N. gruberi* sequences were submitted to RosettaDock in two fragments. Although it is possible that putative *N. gruberi* vaults are hetero-oligomeric, as found for dictyostelids, the RosettaDock modeling indicates that the shorter D2W0Z9 monomers dock along their entire length more readily than do either D2V5B9 or combinations of both (table 2, combination data not shown). Although the energy score graphs do not demonstrate a clear funnel, and therefore less consensus among the models generated by RosettaDock, the energy scores of the docked models are similar to those of the positive control models.

Constraining the *N. gruberi* sequences (D2V5B9 and D2W0Z9) in I-TASSER by 2ZUO*b reduced the models propensity to dock in RosettaDock. Constrained D2V5B9 models were identified by their RMSD—which could be predicted as we knew their starting distance apart, rather than by their energy score, as the energy scores were very similar and the consensus poor. We know from the control studies that the constraint can adversely affect the monomer docking depending on the sequence divergence between the query and constraint structure. It may be that if a constraint needs to be used for very remote sequences such as those found via PSI-BLAST, it would be an improvement to use a high confidence I-TASSER output model from a more closely related species as a constraint in preference over a structure from the PDB. The poorer docking of the more remote *N. gruberi* putative MVP sequences is likely due to the greater divergence of sequence and structure resulting in inaccuracies in monomer modeling. For example, in D2VSY6 (fig. 7*D*), the interruption to the helical structure within the cap-helix section is hindering docking (table 2 and fig. 7). The failure of the rat constraint to improve modeling also reflects this divergence from mammalian MVP sequences.

Given all these results, we propose that *N. gruberi* is capable of making a vault complex with either D2V5B9 or D2W0Z9, both genes have recently been provisionally (and independently of ourselves) reannotated as *mvp* (05/16/12) with the repeat areas additionally annotated as such, thus supporting our results. When used as the query sequence in a UniProt:KB BLAST at default settings, these sequences identify all known MVP sequences. The I-TASSER *C* scores indicate high confidence that the modeled MVP folds are correct and the predicted structures dock along their entire length in RosettaDock. The more remote sequences from *N. gruberi* (DZUF7, D2VSY6, D2VC38, and D2VH38) appear unlikely to be genuine MVP homologs or have diverged significantly from an ancestral MVP sequence. None of DZUF7, D2VSY6, D2VC38, and D2VH38 retrieves any MVP sequences when used as the query sequence in a BLAST at default settings, and although there is some evidence of lateral docking between monomers, this is most likely due to a natural tendency for coils to interact, and the docking does not extend over the entire length as is required for vault formation.

Excavate databases were searched using PSI-BLASTs independently to retrieve sequences with even the slightest resemblance to MVP. Putative MVP sequences from the parasites *Giardia intestinalis* (UniProtKB:C6LY21) and *Trichomonas vaginalis* (UniProtKB:A2FTW3) were also retrieved, but I-TASSER did not identify any kind of convincing MVP structural homolog (data not shown). Interestingly though, a BLAST search with the *G. intestinalis* putative MVP sequence retrieves MVP from both rat and cow within default parameters (*E* values: 9.3 and 4.2, respectively). Additionally, excavate genome databases were searched using the gene sequences from *L. major* and *T. cruzi* without resulting in any hits other than in trypanosomes and leishmanias where in excess of 50 sequences were retrieved. It has been suggested that the trypanosomes evolved from within the bodonids (euglenozoa) (Deschamps et al. 2011). The *Bodo saltans* annotation is incomplete, but if an MVP homolog exists, we should have expected to retrieve something of it. The lack of any readily identifiable putative MVP homolog in any other excavate, based on currently available sequences, is very intriguing. We therefore conclude that even though some protein sequence homology exists within other excavates, our 3D studies indicate that there is no current evidence that other sequenced excavates are capable of forming a vault particle.

Discussion

Three-Dimensional Methodology

The approach described here, using protein structure modeling and docking algorithms, was developed to help answer the question as to the extent that tertiary and quaternary structures will aid the identification of homologous proteins.

The particular application is the question in which species do we find genuine MVP, and if we do, will the MVP monomers form a vault? In this case, BLASTing provides valuable data on the presence of MVP homologs but does not inform directly on the likelihood of any identified MVP monomers assembling into vaults. In general, we need to use more comprehensive methods to demonstrate that limited sequence identity does not preclude vault formation. Here, we show that both tertiary and quaternary structures can be used in addition to information from primary sequences.

It could be argued that the sequence similarity is sufficient for protein prediction servers to be biased toward presenting a structure that is more similar to MVP because there are insufficient alternative templates. However, there are a number of solved structures that could reasonably be ascribed to these sequences, for example, TolA, the stomatin core, band 7 proteins, flotillin, and the colicin membrane spanning protein identified by Phyre[2]. These may hint at possible ancestry for MVP though all are bacterial proteins. Searching for vault specific domains, for example, shoulder or repeats in Pfam (Finn et al. 2010) results in far fewer putative homologs than the BLAST searches. This is undoubtedly because annotation lags far behind sequencing.

It may also be argued that once the structure of the protein is predicted to be MVP-like, then RosettaDock is more likely to find that it does dock. In fact, coiled-coil motifs are likely to dock though usually through twisted supercoiling (Burkhard et al. 2001) rather than lateral association. The I-TASSER-predicted myosin1A coil motifs are docked by RosettaDock, although this example is oversimplified by the absence of the light chains normally present in vivo. However, we have shown for the newly identified MVPs that the lateral docking extends to the shoulder and repeat sections with energy score not dissimilar to the rat and sea urchin where vaults have been observed to form. In the repeat areas in particular, MVP sequence homology is less than 20% versus rat, and our argument is that only those residues that are essential to maintain the shape and lateral docking have been retained.

Although sequence homology of more than 50% is often predictive of structural homology (Clark et al. 2009; Sawyer et al. 2009), there are instances when structure can be dissimilar even with high sequence homology, for example, the prion protein (Pan et al. 1993) and engineered examples (Gronenborn et al. 1991). In this study, we are looking toward the opposite end of the similarity scale, how slim the sequence homology can be and yet structural similarity "sufficient for function" be retained (Holm and Sander 1997). We use MVP as an example to show that structural prediction analysis can extend sequence homology searches. The principles established here could apply to any protein structure. It is more time consuming to check proposed homologies using structural forms but is readily attainable. An important point is that we should not specify too narrow an assumption of the expected structure of a protein. For example, using the rat tertiary MVP structure as a constraint appears to hinder the detection of related structures in the very distantly related excavates and can disrupt docking by RosettaDock.

Seeking traditional homologous sequences through BLAST searches takes just a matter of minutes, with PSI-BLAST a little longer. This is partly why the simple BLAST solution is so attractive. However, methods that test whether sequence homology implies similarity of function, using structural approaches that can detect more distantly related homologs, are more computationally expensive. In general, a protein the length of MVP (~860 residues) is estimated by I-TASSER to take 50 h and is limited to one job per IP address. Both LOMETS and Phyre[2] are very much quicker taking a matter of hours but do not give quantitative results such as the *C* score. LOMETS is limited to one job, but Phyre[2] will accept batch jobs. FATCAT is almost instantaneous, but the RosettaDock server also takes up to 50 h for the 600-residue MVP sections depending on server load. In summary, this is a much slower method than simply BLASTing, but as annotation lags far behind sequencing, we need to go beyond BLASTing and be much more rigorous in our determination of protein homology.

Informing Evolutionary Studies

The evolutionary history of the vault MVP should help identify possible past functions and illuminate current thoughts on function. The big picture questions are these: are vaults ancestral, having been retained in some species, but fallen into disrepair or lost beyond all recognition in others, or alternatively have they been comprised parts that had other functions and have come together in a fairly remote eukaryote and vaults formed thereafter? If we could be confident which species have functional vaults, and which do not appear to have need for them, or possibly maintain the MVP monomer for another purpose, we should be able to clarify their role. We can suggest that this exquisite example of form, with no known fundamental function, was in LECA and as putative MVP has also been reported in bacterial genomes (H6L4P8 provisional annotation MVP) could conceivably have been present in the last universal common ancestor LUCA. It seems unlikely that vaults would be present in some very diverse groups (such as kinetoplasts, alveolates, amoebozoa, and metazoans) but not be present in others. Finding a link between species that do not appear to have a need to maintain the vault and whether vtRNA is associated with it might illuminate an underlying basic function. Equipped with a personal computer, an internet connection, and a means of viewing pdb files, anyone can extend sequence homology analysis to investigation in three dimensions, and we suggest that in silico analysis should routinely be used to check for presumptive structure relationships between potentially ancestrally related proteins. However, that is the work for the future. In all these studies, we require the power from tertiary and quaternary studies to combine with the power of purely sequence-based

studies to enrich the techniques available for molecular evolution.

Supplementary Material

Supplementary materials S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

T.K.D. performed the analyses and wrote the drafts, A.J.S.-S. assisted with the analysis programs, and D.P. designed the original research project. All authors regularly discussed the results and contributed to the final manuscript. This work was supported by internal grants from the Institute of Fundamental Sciences, Massey University.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Berger W, Steiner E, Grusch M, Elbling L, Micksche M. 2009. Vaults and the major vault protein: novel roles in signal pathway regulation and immunity. Cell Mol Life Sci. 66:43–61.

Burkhard P, Stetefeld J, Strelkov SV. 2001. Coiled coils: a highly versatile protein folding motif. Trends Cell Biol. 11:82–88.

Clark AR, Sawyer GM, Robertson SP, Sutherland-Smith AJ. 2009. Skeletal dysplasias due to filamin A mutations result from a gain-of-function mechanism distinct from allelic neurological disorders. Hum Mol Genet. 18:4791–4800.

Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. Mol Biol Evol. 22:1053–1066.

Deschamps P, et al. 2011. Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids. Mol Biol Evol. 28:53–58.

Finn RD, et al. 2010. The Pfam protein families database. Nucleic Acids Res. 38:D211–D222.

Finnigan GC, Hanson-Smith V, Stevens TH, Thornton JW. 2012. Evolution of increased complexity in a molecular machine. Nature 481:360–364.

Fritz-Laylin LK, et al. 2010. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. Cell 140:631–642.

Gray JJ, et al. 2003. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol. 331:281–299.

Gronenborn AM, et al. 1991. A novel highly stable fold of the immunoglobulin binding domain of streptococcal protein-G. Science 253: 657–661.

Hamill DR, Suprenant KA. 1997. Characterization of the sea urchin major vault protein: a possible role for vault ribonucleoprotein particles in nucleocytoplasmic transport. Dev Biol. 190:117–128.

Herrmann C, Zimmermann H, Volknandt W. 1997. Analysis of a cDNA encoding the major vault protein from the electric ray *Discopyge ommata*. Gene 188:85–90.

Holm L, Sander C. 1997. An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. Proteins 28:72–82.

Kedersha NL, Heuser JE, Chugani DC, Rome LH. 1991. Vaults. III. Vault ribonucleoprotein particles open into flower-like structures with octagonal symmetry. J Cell Biol. 112:225–235.

Kedersha NL, Rome LH. 1986. Isolation and characterization of a novel ribonucleoprotein particle—large structures contain a single species of small RNA. J Cell Biol. 103:699–709.

Kelley LA, Sternberg MJ. 2009. Protein structure prediction on the web: a case study using the Phyre server. Nat Protoc. 4:363–371.

Kickhoefer VA, Rome LH. 1994. The sequence of a cDNA encoding the major vault protein from *Rattus norvegicus*. Gene 151:257–260.

Kickhoefer VA, et al. 2009. Targeting vault nanoparticles to specific cell surface receptors. ACS Nano 3:27–36.

Koonin EV. 2010. The incredible expanding ancestor of eukaryotes. Cell 140:606–608.

Kozakov D, Brenke R, Comeau SR, Vajda S. 2006. PIPER: an FFT-based protein docking program with pairwise potentials. Proteins 65: 392–406.

Kozlov G, et al. 2006. Solution structure of a two-repeat fragment of major vault protein. J Mol Biol. 356:444–452.

Kurland CG, Collins LJ, Penny D. 2006. Genomics and the irreducible nature of eukaryote cells. Science 312:1011–1014.

Kuwahara Y, et al. 2009. Unusual thermal disassembly of the SPFH domain oligomer from *Pyrococcus horikoshii*. Biophys J. 97:2034–2043.

Lara PC, Pruschy M, Zimmermann M, Henriquez-Hernandez LA. 2011. MVP and vaults: a role in the radiation response. Radiat Oncol. 6: 148.

Liu B, et al. 2011. Up-regulation of major vault protein in the frontal cortex of patients with intractable frontal lobe epilepsy. J Neurol Sci. 308: 88–93.

Lyskov S, Gray JJ. 2008. The RosettaDock server for local protein-protein docking. Nucleic Acids Res. 36:233–238.

Matsumoto T, et al. 2011. Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. Plant Physiol. 156:20–28.

Mossel E, Steel M. 2004. A phase transition for a random cluster model on phylogenetic trees. Math Biosci. 187:189–203.

Moult J, Pedersen JT, Judson R, Fidelis K. 1995. A large-scale experiment to assess protein structure prediction methods. Proteins 23:ii–iv.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP—a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 247:536–540.

Neumann N, Lundin D, Poole AM. 2010. Comparative genomic evidence for a complete nuclear pore complex in the last eukaryotic common ancestor. PloS One 5:e13241.

Orengo CA, et al. 1997. CATH—a hierarchic classification of protein domain structures. Structure 5:1093–1108.

Pan KM, et al. 1993. Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins. Proc Natl Acad Sci U S A. 90:10962–10966.

Paspalas CD, et al. 2009. Major vault protein is expressed along the nucleus-neurite axis and associates with mRNAs in cortical neurons. Cereb Cortex. 19:1666–1677.

Qian W, He X, Chan E, Xu H, Zhang J. 2011. Measuring the evolutionary rate of protein-protein interaction. Proc Natl Acad Sci U S A. 108: 8725–8730.

Querol-Audi J, et al. 2009. The mechanism of vault opening from the high resolution structure of the N-terminal repeats of MVP. EMBO J. 28: 3450–3457.

Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 5: 725–738.

Roy A, Xu D, Poisson J, Zhang Y. 2011. A protocol for computer-based protein structure and function prediction. J Vis Exp. 57:e3259.

Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 234:779–815.

Sawyer GM, Clark AR, Robertson SP, Sutherland-Smith AJ. 2009. Disease-associated substitutions in the filamin B actin binding domain confer enhanced actin binding affinity in the absence of major structural disturbance: insights from the crystal structures of filamin B actin binding domains. J Mol Biol. 390:1030–1047.

Schäffer AA, et al. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29:2994–3005.
Stadler PF, et al. 2009. Evolution of vault RNAs. Mol Biol Evol. 26: 1975–1991.
Steitz TA, Moore PB. 2003. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. Trends Biochem Sci. 28:411–418.
Stephen AG, et al. 2001. Assembly of vault-like particles in insect cells expressing only the major vault protein. J Biol Chem. 276: 23217–23220.
Stevens MI, Hunger SA, Hills SFK, Gemmill CEC. 2007. Phantom hitch-hikers mislead estimates of genetic variation in Antarctic mosses. Plant Systematics Evol. 263:191–201.
Stewart PL, et al. 2005. Sea urchin vault structure, composition, and differential localization during development. BMC Dev Biol. 5:3.
Tanaka H, et al. 2009. The structure of rat liver vault at 3.5 Angstrom resolution. Science 323:384–388.
Tovchigrechko A, Vakser IA. 2006. GRAMM-X public web server for protein-protein docking. Nucleic Acids Res. 34:W310–W314.
van Zon A, et al. 2002. Structural domains of vault proteins: a role for the coiled coil domain in vault assembly. Biochem Biophys Res Commun. 291:535–541.
Vasu SK, Kedersha NL, Rome LH. 1993. cDNA cloning and disruption of the major vault protein alpha gene (mvpA) in *Dictyostelium discoideum*. J Biol Chem. 268:15356–15360.
Vasu SK, Rome LH. 1995. Dictyostelium vaults: disruption of the major proteins reveals growth and morphological defects and uncovers a new associated protein. J Biol Chem. 270:16588–16594.
Vollmar F, et al. 2009. Assembly of nuclear pore complexes mediated by major vault protein. J Cell Sci. 122:780–786.
Watson JD, Laskowski RA, Thornton JM. 2005. Predicting protein function from sequence and structural data. Curr Opin Struct Biol. 15:275–284.
Wu S, Zhang Y. 2008. MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72:547–556.
Xu D, Zhang J, Roy A, Zhang Y. 2011. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins 79: 147–160.
Yang J, et al. 2010. Vaults are dynamically unconstrained cytoplasmic nanoparticles capable of half vault exchange. ACS Nano 4: 7229–7240.
Ye Y, Godzik A. 2003. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics 19:ii246–ii255.
Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:409.

Associate editor: Dan Graur