Research

Comprehensive *de novo* structure prediction in a systems-biology context for the archaea *Halobacterium* sp. *NRC-1*

Richard Bonneau, Nitin S Baliga, Eric W Deutsch, Paul Shannon and Leroy Hood

Address: Institute for Systems Biology, Seattle, WA 98103-8904, USA.

Correspondence: Leroy Hood. E-mail: lhood@systemsbiology.org

Published: 12 July 2004

Genome ***Biology*** 2004, **5:**R52

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/8/R52

Received: 5 March 2004
Revised: 7 March 2004
Accepted: 1 June 2004

Abstract

Background: Large fractions of all fully sequenced genomes code for proteins of unknown function. Annotating these proteins of unknown function remains a critical bottleneck for systems biology and is crucial to understanding the biological relevance of genome-wide changes in mRNA and protein expression, protein-protein and protein-DNA interactions. The work reported here demonstrates that *de novo* structure prediction is now a viable option for providing general function information for many proteins of unknown function.

Results: We have used Rosetta *de novo* structure prediction to predict three-dimensional structures for 1,185 proteins and protein domains (<150 residues in length) found in *Halobacterium NRC-1*, a widely studied halophilic archaeon. Predicted structures were searched against the Protein Data Bank to identify fold similarities and extrapolate putative functions. They were analyzed in the context of a predicted association network composed of several sources of functional associations such as: predicted protein interactions, predicted operons, phylogenetic profile similarity and domain fusion. To illustrate this approach, we highlight three cases where our combined procedure has provided novel insights into our understanding of chemotaxis, possible prophage remnants in *Halobacterium NRC-1* and archaeal transcriptional regulators.

Conclusions: Simultaneous analysis of the association network, coordinated mRNA level changes in microarray experiments and genome-wide structure prediction has allowed us to glean significant biological insights into the roles of several *Halobacterium NRC-1* proteins of previously unknown function, and significantly reduce the number of proteins encoded in the genome of this haloarchaeon for which no annotation is available.

Background

The archaeon *Halobacterium NRC-1* is an extreme halophile that thrives in saturated brine environments such as the Dead Sea and solar salterns. It offers a versatile and easily assayed system for an array of well-coordinated physiologies that are necessary for survival in its harsh environment [1]. It has robust DNA repair systems that can efficiently reverse the damages caused by a variety of mutagens including UV radiation and desiccation/re-hydration cycles [2,3]. *Halobacterium NRC-1* adapts its metabolism to anaerobic conditions with the synthesis of bacterorhodopsin, which facilitates the conversion of energy from light into ATP. The completely

sequenced genome of *Halobacterium NRC-1* (containing ~2,600 genes) has provided insights into many of its physiological capabilities, however nearly half of all genes encoded in the halobacterial genome have no known function [4-7].

This work is intended to be a prototype for the development of a biological data integration system with a focus on identifying putative functional predictions for proteins of unknown function derived from *de novo* protein structure predictions. The main result is a reannotation of the *Halobacterium NRC-1* proteome that includes general functional information gleaned from protein structure prediction that can be explored in the context of the predicted association network for the *Halobacterium NRC-1*. The information is derived using Rosetta *de novo* structure prediction as part of a combined annotation pipeline that also includes several primary-sequence similarity based methods. The annotation pipeline organizes several annotation methods in a hierarchy such that sequence-based methods (such as PSI-BLAST and Pfam) are applied first; we rely on structure prediction as a source of function information only for those proteins not annotated via primary sequence based methods. We illustrate the merits of our approach by bringing three examples to the forefront where integrating Rosetta with one or more independent methods for predicting functional associations, described herein, produces functional conclusions not accessible by any single method.

One paradigm for predicting the function of proteins of unknown function, the so called 'sequence-to-structure-to-structure-to-function' paradigm, is based on the assumption that three-dimensional structure patterns are conserved across a much greater evolutionary distance than recognizable primary sequence patterns [8]. This assumption has been supported by several structure-function surveys of the Protein Data Bank (PDB) which show that fold similarities in the absence of sequence similarities imply some shared function in the majority of cases [9-13]. One protocol for predicting protein function based on this paradigm is to predict the structure of a query sequence of interest and then use the predicted structure to search for fold or structural similarities between the predicted protein structure and experimentally determined protein structures in the PDB or a non-redundant subset of the PDB [14-17]. There are, however, several problems associated with deriving functional annotation from fold similarity - fold similarities can occur through convergent evolution, and thus have no functional implications. Also, aspects of function, such as precise ligand specificity, can change throughout evolution leaving only general function intact across a given fold superfamily [18-20]. Fold matches between the predicted structures and the PDB are thus treated as sources of putative general functional information and are functionally interpreted primarily in combination with other methods such as global expression analysis and the predicted protein association network. In this study we use Rosetta to generate a confidence ranked list of possible structures for proteins and protein domains of unknown function, search each of the ranked structure predictions against the PDB, and then calculate confidences for the fold predictions and evaluate possible functional roles in the context of the Halobacterium association network.

Rosetta is a computer program for *de novo* protein structure prediction, where *de novo* implies modeling in the absence of detectable sequence similarity to a previously determined three-dimensional protein structure [21,22]. Rosetta uses information from the PDB to estimate possible conformations for local sequence segments (three and nine residue segments). It then assembles these pre-computed local structure fragments by minimizing a global scoring function that favors hydrophobic burial and packing, strand pairing, compactness and energetically favorable residue pairings. Results from the fourth and fifth critical assessment of structure prediction (CASP4, CASP5) have shown that Rosetta is currently one of the best methods for *de novo* protein structure prediction and distant fold recognition [23-27]. Using Rosetta generated structure predictions we were previously able to recapitulate or predict many functional insights not detectable from primary sequence [28,29]. Rosetta was also recently used to generate both fold and function predictions for Pfam [30,31] protein families that had no link to a known structure, resulting in ~120 high confidence fold predictions. In spite of these successes, Rosetta has a significant error rate, as do all methods for distant fold recognition and *de novo* structure prediction. The Rosetta confidence function partially mitigates this error rate by assessing the accuracy of predicted folds [29]. Another unavoidable source of uncertainty, with respect to function prediction, is the error associated with distilling function from fold matches described above. The predictions generated by *de novo* structure prediction are thus best used in combination with other sources of putative or general functional information such as proximity in protein association or gene regulatory networks.

We separate annotations into two classes. The first type are annotations referring to individual proteins, such as structure predictions and function annotations derived from sequence or structural similarity to a protein of known function. The second type are annotations referring to the context of a protein relative to other proteins within the proteome, grouping multiple genes into an operon, two proteins having a similar phylogenetic profile - implying that they carry out related functions, correlation of mRNA or protein concentrations across a variety of genetic or environmental perturbations, and so on. The major difference between association/contextual information, (information of the second type above) and individual annotations (first type) is brought sharply into focus when one considers that there are several highly connected clusters of proteins in the Halobacterium association network that are composed entirely of genes of unknown function. Thus, contextual information (often presented in the form of protein interaction or association networks) must

be combined with methods for extending our ability to infer putative functions for proteins of unknown function if our goal is to understand biological systems globally. In this work contextual annotations have been reduced to pairwise relationships between proteins and are represented graphically as edges of different types in what we will refer to as the Halobacterium association network. Our association network is composed of operon predictions, conserved chromosomal proximity relationships, phylogenetic profile similarity, the occurrence of two *Halobacterium* proteins fused into a single protein in other genomes (domain fusion), and predicted protein-protein interactions primarily derived from yeast two-hybrid studies and *Helicobacter pylori* protein-protein interaction studies. The results of previously described genome-wide mRNA and protein expression studies are mapped onto this network [32,33]. The generation and visualization of this association network is described fully in Materials and methods.

Halobacterium NRC-1 uses an increase in the surface negative charge of its proteins as a major adaptation to a high salt environment. The average protein domain (~150 amino acids) in *Halobacterium* has a net charge of -17 in contrast to -3 for *Saccharomyces cerevisiae* [34]. A higher overall surface charge (and thus fewer surface hydrophobics) will probably reduce a source of error in Rosetta *de novo* predicted conformations - incorrect burial of surface hydrophobics. Additionally, it has been shown that archaeal proteins have shorter loops and that, when present, cysteines are found paired in disulphide bridges more often than in their corresponding eukaryotic homologs [35,36]. These considerations should result in more accurate Rosetta predictions for these halophilic archaeal proteins.

A multi-institutional effort is currently underway to study the genome-wide response of *Halobacterium NRC-1* to its environment. This systems biology effort elevates the need for applying improved methods for annotating proteins of unknown function found in the *Halobacterium NRC-1* genome. Genome-wide measurements of mRNA transcripts, protein concentrations, protein-protein interactions and protein-DNA interactions generate rich sources of data on proteins - those with both known and unknown functions [3,32]. Often these systems-level measurements do not suggest a unique function for a given protein of interest, but instead suggest their association with, or perhaps their direct participation in, a previously known cellular function. Thus, investigators using genome-wide experimental techniques are now routinely generating data for proteins of hitherto unknown function that appear to play pivotal roles in their studies. Proteins of partially known function can also present challenges to methods for function assignment, as many of these proteins have large regions of sequence of unknown function - that is, many proteins have multiple domains only one (or a few) of which are homologous to proteins of known function. These mystery-proteins and mystery-domains require the development of computational methods that can be used to better determine functional roles for proteins and protein-domains of unknown function.

Results and discussion

Structure prediction

We have applied our annotation pipeline to 2,596 predicted proteins in the *Halobacterium NRC-1* genome (Figure 1). This pipeline represents an annotation hierarchy wherein for each protein we first attempt function assignment on the basis of primary sequence similarity to characterized proteins or protein families; this step includes algorithms such as PSI-BLAST and HMMER searches, both of which have low false positive rates and well characterized error models. In instances where primary sequence similarity methods fail to assign putative functions to the proteins, we predict their three-dimensional structures primarily using two methods: Rosetta *de novo* structure prediction and Meta-Server/3D-jury fold recognition. Rosetta structure prediction is only applicable to proteins and protein domains fewer than 150 residues in length and thus separating proteins into domains prior to analysis is key to the success of our approach. We have used Ginzu, a program that detects proteins domains boundaries using Pfam and PSI-BLAST alignments, to separate proteins into domains prior to annotation [37]. This resulted in 1,926 proteins containing a single domain and 670 proteins that could be divided into 1,665 domains (a total of 3,591 proteins and protein-domains were analyzed by the annotation pipeline). These 3,591 domains included both proteins of known function, annotated as part of the initial annotation [7], and proteins that were unannotated at the time this study began.

Of the 2,596 proteins in the *Halobacterium* genome, 1,077 had significant matches by PSI-BLAST to known structures in the PDB. An additional 610 domains lacking PSI-BLAST hits to the PDB had matches to Pfam protein families (detected using HMMER). Following the application of the above methods, Rosetta was used to predict the three-dimensional structures of all proteins and protein domains (<150 residues in length) for which we were unable to detect significant matches to known structures or Pfam domains. The application of the Rosetta method to these domains of unknown function resulted in 670 high confidence fold predictions that were then analyzed in the context of the *Halobacterium NRC-1* association network. A total of 1,234 proteins eluded all attempts to predict fold and/or function using any of the methods described above: 239 domains were too large for Rosetta, and 995 domains produced low confidence Rosetta results. The results of this annotation hierarchy are publicly available on the Institute for Systems Biology (ISB) Halobacterium Research Resource website [38]. Using this approach we were able to significantly reduce the number of protein domains in the *Halobacterium NRC-1* genome for which no hits by any method were found.

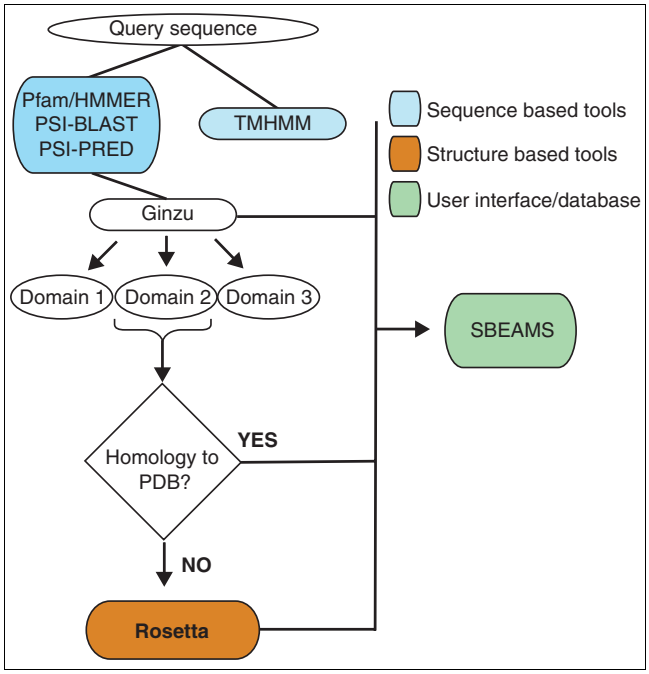

Figure 1
Flow chart depicting the annotation pipeline implemented in this study. Sequence based methods are employed first (top), domains that elude primary sequence based methods are predicted by structure-prediction methods (bottom). For any given genome, data from all levels in this method hierarchy are integrated using SBEAMS (Systems Biology Experiment Analysis and Management System). Implicit in this annotation hierarchy is the idea that protein annotation should be domain-centric (that is, multi-domain proteins should be divided into domains as early as possible in the annotation process). SBEAMS produces a table of annotations where for a given domain only results from the topmost level in the method hierarchy (PDB-BLAST → Pfam → Rosetta) producing a significant hit are displayed.

Association network construction

We have constructed a network of predicted protein associations composed of several different edge types as described below.

We identified pairs of interacting orthologs using the database of Clusters of Orthologous Genes (COGs) [39] in combination with databases of protein interactions in other organisms (*S. cerevisiae*, *H. pylori*) [40-43]. Putative protein binding pairs in *Halobacterium* sp. were inferred in three stages: COG members of the protein-protein interaction pairs were determined; all corresponding COG orthologs of yeast and *H. pylori* interacting proteins were identified in the *Halobacterium* sp. genome; and the interacting pairs were given a confidence level determined by the strength of the match to the COG pair and the confidence of the original protein interaction measurements. A total of 1,143 non-redundant predicted interactions were inferred by this method.

We use the method of Marcotte and colleagues [44,45], commonly referred to as the phylogenetic profile method, to detect groups of genes with significant co-occurrence across multiple genomes. This added 525 phylogenetic profile edges to the *Halobacterium NRC-1* association network. Enright *et al.* [46] have demonstrated that domain fusions are often correlated with functional interactions among the corresponding domains. Domains within *Halobacterium* sp. proteins fused in other genomes have been used as a metric to predict functional associations between the corresponding proteins. We identified 2,460 putative associations with this approach. These putative functional couplings include: proteins that may participate in the same biochemical pathway, proteins that may interact with each other, and/or proteins that may be co-regulated in response to a common environmental stimulus.

Two methods were used to detect significant co-localization of genes into operons and evolutionarily conserved chromosomal proximities. These groupings were then represented as pairwise interactions in the association network. The most reliable method of the two methods employed requires two proximal genes to have homologs (via the COG database) in close proximity in at least one other genome [47] - 327 such conserved operons pairs were detected and added to the association network. This conserved proximity method requires a pair of genes to have orthologs in other genomes and is thus not applicable for genes lacking homologs in other genomes. Another method for operon prediction [48], which does not require genes to be found in other systems, relies only on the proximity and co-directionality of genes, that is, that genes be on the same strand and close compared to the distribution of chromosomal proximity for the genome as a whole. A total of 1,335 statistically significant proximities were added to the network as potential operon edges.

Integration of the different edge types and the experimental data (primarily microarray expression data) with the annotation table was carried out using Cytoscape [49] and the Systems Biology Experiment Analysis and Management System (SBEAMS). SBEAMS is a modular framework for collecting, storing, accessing and integrating data produced by various experiments using a relational database that is being actively developed and maintained at the ISB. Cytoscape is a network visualization tool that allows for the simultaneous viewing of biological networks with several types of biological data, such as global mRNA expression data. All microarray data used in this study are the result of previously described studies [3,32]. The association network is available as a Java-web-start on the ISB Halobacterium Research Resource website [38].

Highlights

Although the results of this analysis are publicly available via our website we will briefly outline three cases where the Rosetta data, along with the association network, were useful in annotating proteins of previously unknown function and thus furthered our understanding of *Halobacterium NRC-1*

biology, by generating a biologically relevant and testable hypothesis.

Insight into chemotaxis

Halobacterium NRC-1 can physically relocate to favorable environments by virtue of a bacterial-like chemotaxis system (chemotaxis, flagellar motor, and several signaling Htr/methyl-accepting chemotaxis genes). The chemotaxis system receives signals from sensors for light (Sop1-Htr1 and Sop2-Htr2), oxygen (Htr8) [50], amino acids and sugars as well as a variety of other small molecules [51]. Briefly, signals from the environment are received by a sensing domain and transmitted to a methyl-accepting signaling domain (for example, Htr1/Htr2). Htr proteins then transmit this signal to the flagellar machinery via CheA. Here we have defined Htr proteins as those proteins containing the methyl-accepting chemotaxis domain, MCP, and sometimes containing a HAMP domain (Pfam doamain PF00672) (HAMP domains are often associated with MCP domains and they are found to be essential in transmitting signals between sensory input modules and MCP domains) [52-54]. The *Halobacterium NRC-1* genome encodes 17 such Htr proteins; Figure 2 shows these 17 Htr proteins and their neighbors in the association network. All Htr proteins are connected to CheA via phylogenetic profile edges, probably reflecting the fact that Htr methyl-accepting chemotaxis proteins are known to physically interact with CheA. In general there is a much greater diversity in the sensing domains than in the Htr domains they interact with, and, thus, we have yet to identify the corresponding sensing domain for several of the Htr proteins in the *Halobacterium NRC-1* genome. Htr domains and their corresponding sensing domains are found both encoded on the same polypeptide as well as on separate polypeptides with co-regulated expression levels. For instance, Sop1 and Sop2 are co-transcribed with their signal transducers, Htr1 and Htr2 respectively, as separate proteins organized into two-protein operons (see Figure 2). Alternatively, Htr8 is transcribed as a single two-domain protein containing both a sensing domain and the Htr domain on a single polypeptide chain. In five cases (Htr1, Htr2, Htr3, Htr5, Htr8) the function (with respect to sensing specificity) was known prior to this study. In the case of *htr5* the identity of its operon complement (*proX*) points to roles in chemotaxis towards osmoprotectants and amino acids [55]. Htr3 has also been shown to be responsible for chemotactic response towards leucine, isoleucine, valine, methionine and cysteine. Along similar lines, *htr18* has an operon edge to *potD*, suggesting chemotaxis in response to lipids (Table 1), although this has yet to be experimentally verified. For the remaining Htr proteins, for which the corresponding sensing domain has not been identified, we first examine unannotated (non-MCP, non-HAMP) domains in the same polypeptide and subsequently examine unannotated proteins found in the same operon as the Htr in question, as these are the domains in the genome most likely to function as the required cognate sensing domain.

Domains in the Htrs *Htr9*, *Htr14* and *Htr16*, other than MCP and HAMP domains, had matches to Pfam families. Htr9 contains a PAS domain suggesting a role in mediating responses to oxygen or redox-potential (see Table 1). Htr14 had weak matches to the KE2 domain as well as Prefoldin. The co-occurrence of these two domains seems plausible, as Prefoldin and KE2 domains are known to physically interact, although the functional implication of the similarity to these long helical domains is limited. The second domain of Htr16 had a match to apolipoprotein A1 (PF01442), a long domain involved in the uptake of cholesterol and lipids. Again, the functional implications of matches to this long helical (coiled-coil) domain are limited.

In the case of Htr6, its operon complement YufN was annotated as an ABC transporter without any ligand specificity. YufN, however, has a strong match when analyzed by MetaServer/3D-jury to a D-glactose/D-glucose periplasmic binding protein, suggesting that Htr6 and YufN sense sugar in the environment. This is supported by the observation that *Halobacterium NRC-1* has a chemotaxis response towards glucose [51].

The Rosetta-predicted structure for the domain of unknown function in Htr10 matched hemoglobin (PDB: 1ljwA) suggesting a role in aerotaxis and/or redox sensing. In the time elapsed since the Rosetta structure prediction was generated, the crystal structure of the Hemat sensor domain from *B. subtilis* (1OR4) was deposited in the PDB; we can now find a match between domain 1 of Htr10 and 1OR4-A via PSI-BLAST. As shown in Figure 2, our predicted structure match to 1LJW-A is validated by this newly detected sequence match to 1OR4-A. Htr12 has a amino-terminal Pfam match to PF00989 (PAS domain). 3D-Jury also detects (via FFAS) a hit to 1D06 (oxygen-sensing domain of *Rhizobium meliloti* Fixl) over this amino-terminal region of Htr12. VNG1440h, connected to Htr12 by operon edge, had a separate Rosetta-predicted structure match to cytochrome C (2occC2). These findings suggest a role for VNG1440h and Htr12 in aerotaxis sensing. For Htr15, 3D-jury detects a hit to 1DP6-A (also the heme binding domain of FixL). Given that changes in salinity can dramatically alter the solubility of oxygen, the natural evironment of *Halobacterium NRC-1* is in a perpetually dynamic state with respect to oxygen.

With a combination of structure prediction and sequence-based methods we were able gain insights into, and suggest testable hypothesis for, all but three of the 17 proteins with recognizable Htr domains; for Htr4, Htr7 and Htr17 we have found no significant hits by any of the described methods.

Support for the viral origin of the minichromosomes

Prophage regions have been found in the majority of prokaryotic genomes and phage integration into prokaryotic genomes, as well as phage-mediated rearrangements of prokaryotic genomes, has become recognized as a major

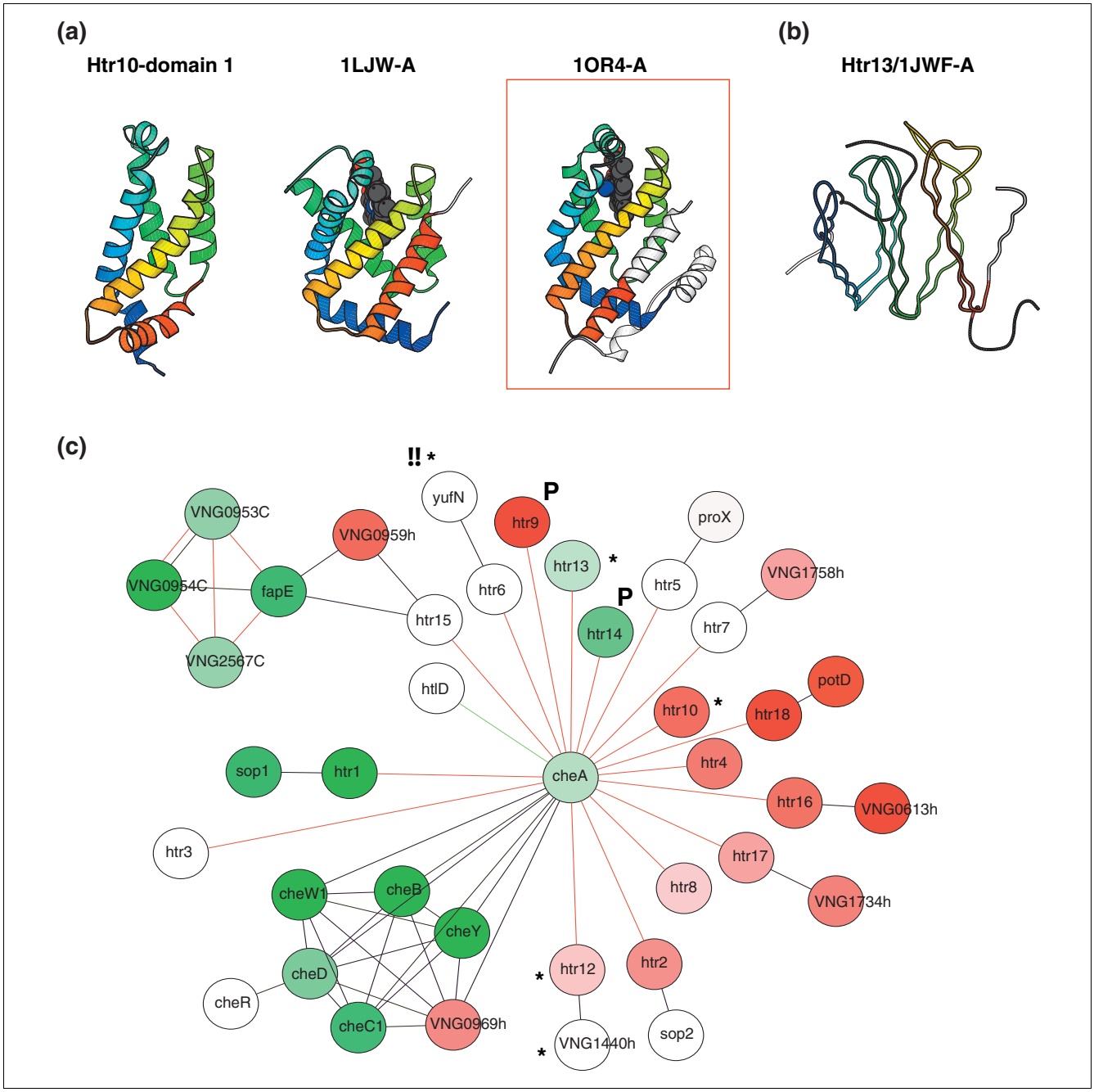

Figure 2
Chemotaxis methyl accepting domains. **(a)** Htr10 (VNG1505g) domain 1 hit to 1ljwA, hemoglobin. The recently deposited structure for the Hemat Sensor domain (1OR4-A) is also shown (red box). The position of the heme (black spheres) is similar in both our predicted fold match (1LJW-A) and the match detected by PSI-BLAST (1OR4-A) **(b)** Htr13 (VNG1013g) hit to Gga1 (1jwfA, involved in protein transport, binding of dipeptide signal sequence), **(c)** the association network surrounding CheA and its interactions with the Htr methyl accepting domains found in the *Halobacterium* genome, as predicted by the phylogenetic profile method (red lines). Also shown are predicted operon edges (black lines). The expression levels (where red corresponds to a high level of expression and green to a low expression relative to a reference; white indicates no change/no measurement) are from a previously described microarray experiment. Nodes marked with asterisks indicate proteins where a domain was folded with Rosetta (resulting in a significant fold match) or annotated using fold recognition. Nodes marked with a 'P' are proteins that were annotated using Pfam. The '!' by yufN indicates that the prior annotation does not agree with our current analysis.

source of lateral gene transfer and a major influence on prokaryotic genome structure [56-59]. Most of this work, however, has been carried out in bacteria-bacteriophage systems (see Table 2). To date, relatively few archaeal viruses have been isolated and studied, although it is known that high counts of phage particles are found in environments dominated by archaea, including hypersaline environments [60,61]. This leads us to believe that phage integration events might have played important roles in the evolution of *Halobacterium NRC-1* through lateral gene transfer and chromosomal rearrangements. The 2.6 Mbp genome of *Halobacterium NRC-1* is organized as a 2,014kb chromosome of 68% G+C and two smaller replicons pNRC100 (191 kb) and pNRC200 (365 kb) of an average of 58% G+C [62,63]. One striking feature of the *Halobacterium NRC-1* genome is the presence of an unusually large number of insertion sequences (IS-elements); remarkably, 69 of the 91 IS-elements localize to the two minichromosomes, which together constitute only 22% of the complete genome. Most of these IS-elements code for transposases, which are often associated with phage genomes and other mobile genetic elements [64,65]. Thus, the high IS-element density combined with the lower G+C content of the two minichromosomes, leads us to believe that IS-element rich regions of the minichromosomes have been recently introduced via lateral gene transfer. These IS-element rich regions have a higher percentage of proteins of unknown function and are thus ideally suited for analysis by the Rosetta method. We have examined several IS-element rich regions of the large (pNRC200) and small (pNRC100) minichromosomes and found evidence that these regions are highly divergent prophage remnants.

The IS-element region on pNRC200 spanning genes VNG6098H to VNG6121H is duplicated on pNRC100 (VNG5101H to VNG5124H). We find several matches within this region to the Pfam transposase DDE (PF01609) family of proteins, first isolated from bacteriophage lambda [66]. We also find two matches to phage integrases: VNG6112H matches Pfam integrase family PF00589, while two Rosetta-predicted structures for VNG6117H both match a phage integrase fold (1asu00). Finally, the Rosetta-predicted structure for VNG6105H matches the capsid protein from Rous sarcoma virus (PDB: 1d1dA2). Thus our analysis suggests that this region is a prophage remnant that has since diverged and may no longer function as an active or complete phage.

Likewise, the region on the small minichromosome from VNG5040H to VNG5051G (duplicated on the reverse strand as VNG5256 to VNG5246G) was also investigated as a region likely to harbor prophage remnants. Again, we see many examples of matches to the transposase DDE domain family (PF01609). This region contains genes encoding a TATA-binding protein (TBP), TbpB, and a second protein (VNG5048) whose Rosetta-predicted structure also matches a TBP fold. For VNG5049H/5248H, we find a Rosetta predicted structure match to 2ezh, a fold that has been observed in both transposases and transcription factors. We also find a match for a Rosetta-predicted structure to an HIV capsid protein (1am3) for VNG5047H (Figure 3e). Thus in this region (VNG5040H to VNG5051G), as in the region described above (VNG6098H to VNG6121H), we have found hits to several proteins often found within phage/viral genomes (transposase, integrase, capsid proteins, general transcription factors) possibly indicating that these regions are highly divergent prophage remnants.

Proposed transcriptional regulators

The detection of potential transcriptional regulators in the *Halobacterium NRC-1* genome is central to an ongoing systems biology effort aimed at understanding its regulatory mechanisms and ultimately its global gene regulatory circuitry. The majority of archaeal transcriptional regulators, detected by sequence similarity, are small helical domains that are well within the size and complexity limit of the Rosetta structure prediction method. Therefore, we have examined similarities of Rosetta-predicted structures to the CATH fold families that correspond to the following archaeal transcriptional regulators: 1.10.10.10, 1.10.10.60 and 1.10.472.10 [16]. Our results are summarized in Table 3 and Figure 4. Although the majority of prokaryotic proteins sharing these folds are involved in transcriptional regulation, in general, a match via sequence or structure-based methods to these small DNA binding domains can also have other functional interpretations. Therefore, an important caveat is that these function-predictions stand as putative transcription regulators or DNA binders. Ongoing analysis of global gene expression patterns in *Halobacterium NRC-1* will probably provide additional complementary information that will help clarify the precise functional roles of these proteins. Many of the predicted DNA binders or regulators described in Table 3 also have weak sequence similarities to other transcriptional regulators of known structure, detected via PSI-BLAST or fold recognition, which support the Rosetta predictions.

We describe below, three cases below where weak matches by PSI-BLAST or FFAS to transcriptional regulators support our *de novo* structure predictions and result in higher confidence function predictions than possible by any one method. Furthermore, since transcriptional repressors and activators are often encoded at termini of the operon the regulate [67], we use functional associations such as conserved chromosomal proximity and predicted operons, to suggest potential targets for these regulators.

In the first case, we find a Rosetta-predicted structural match for VNG0462C to the diphtheria toxin repressor (1bi2-B) and also find a weak similarity by PSI-BLAST to a transcriptional regulator 1lnwA (1lnwA is a fold similar to that of the diphtheria toxin repressor). Furthermore, *VNG0462* is in an operon with *VNG0463H*, a membrane protein of unknown function, and *nosF2*, a copper transport ATPase. Consistent

Table 1

Htr1-Htr18 chemotaxis annotations

Gene name	Name	Sensing domain	Length	HAMP domain	Membrane regions	Method	Role/responds to	Annotation
htr1	VNG1659g	sop1	536	35-104	12-31	Known	Light	Responds to light via sensory rhodopsin
htr2	VNG1765g	sop2	764	283-352	13-35	Known	Light	Responds to light via sensory rhodopsin
htr3	VNG1856g	self/?	633	125-195	125-144	Known	Amino acids	
htr4	VNG0806g	self/?	778	298-367	29-48, 297-319	-	-	
htr5	VNG1760g	ProX	810	325-394	35-57, 325-344	Known	Amino acids, osmoprotectants	ProX is a putative glycine betatine/ choline/proline substrate-binding protein
htr6	VNG0793g	yufN	545	295-365	21-43, 297-319	3D-Jury	Sugars	yufN is annotated as an ABC transporter and lipoprotein META-SERVER/3D-Jury finds a strong hit to 2 gbp D-galactose/D-glucose periplasmic binding protein
htr7	VNG1759g	VNG1758H	789	-	1-91	Rosetta	-	Weak hit to sensory rhodopsin
htr8	VNG1523g	self	633	-	48-206	Known	Oxygen	Experimentally known to play a role in aerotaxis
htr9	VNG1395g	self/?	481	-	-	Pfam	Redox/o2/light	PF0989, PAS domain
htr10	VNG1505g	self/?	489	-	-	Rosetta PSIBLAST	Oxygen	Domain 1 rosetta hit to 1ljwA hemoglobin Domain 1 hit to 1or4A via PSI-BLAST (recent PDB)
htr12	VNG1442g	self/ VNG1440H	420	-	-	Rosetta 3D-jury Pfam	Redox/o2/light	htr12 has amino-terminal (domain 1) hit to PF0989 (PAS domain) htr12-domain 1 also has a 3d-Jury hit to 1dp6A (FixL heme domain) VNG1440H has a weak rosetta hit to 2occC2 (cytochrome C subunit) and predicted TM helices
htr13	VNG1013g	self/?	423	-	-	Rosetta	Peptides/?	Hit to 1jwfA (Gga1, involved in protein transport) Htr13 could be involved in response to peptides in the environment
htr14	VNG0355g	self/?	627	58-129	36-58	Pfam	Peptides/?	Weak hits to PF01920-KE2 domain, PF02996-prefoldin
htr15	VNG0958g	VNG0959H	636	-	-	3D-jury	Oxygen	Domain 2 has 3d-jury hit to 1dp6A (FixL Heme domain)
htr16	VNG0614g	self/ VNG0613H	628	129-199	130-152	Pfam	Lipids/?	Hit to PF01442 in domain 2 of htr16 (PF01442 is an apolipoprotein Involved in the uptake of lipids and/or cholesterol)
htr17	VNG1733g	VNG1734H	536	-	1-91	-	-	Not applicable
htr18	VNG0812g	PotD	790	257-327	-	known	Lipids	PotD is a spermidine/putrescine binding protein

with this organization, the three genes are co-regulated at the mRNA levels in our microarray experiments [3,32]. Therefore, our current hypothesis is that VNG0462C is regulating these two proteins and that the three gene operon is involved in copper transport.

In the second case, the Rosetta predicted structure for VNG5156H also matches the diphtheria toxin repressor (1bi2-B) and is consistent with a weak match by FFAS to a winged helix transcriptional regulator (1i1gA). VNG5156H is in an operon with three genes all encoding proteins of unknown function. One of these proteins, VNG5154H, has a carboxy-terminal domain that matches the restriction endonuclease domain family PF04471. Again, we observe that the genes encoded in this operon are highly co-regulated at the mRNA level across all of our microarray experiments. Thus, one likely hypothesis is that this operon of unknown function is regulated by VNG5156H, but the exact functions

Table 2

Annotations for IS element rich regions

Name	IS-element	Method	Annotation
IS-element rich region 1			
VNG5101H/6098H	ISH2	Pfam	PF01402 CopG ribbon helix, regulates plasmid copy number
VNG5102H/6099H	-	TMHMM	Membrane protein, unknown function
VNG5104H/6101H	-	-	-
VNG5105H/6102H	-	Meta-Server	Hit to 1dhx, Coper binding protein
VNG5106H/6103H	-	TMHMM	Membrane protein, unknown function
VNG5108H/6105H	-	Rosetta	Hit to 1d1A2, capsid protein/transcription factor
VNG5109H/6106H	ISH8	Pfam	PF01609, transposase DDE domain
VNG5112H/6109H	-	Rosetta	Hit to 1dt9A1, translation initiation factor
VNG5114H/6111H	ISH3	-	-
VNG5115H/6112H	-	Pfam	PF00589, phage integrase family
VNG5116H/6113H	-	Meta-Server	Hit to 1d1qA, phosphotyrosine protein phosphatase
VNG5118H/6115H	-	Rosetta	1he8A3, serine/threonine protein phosphatase
VNG5119H/6116H	-	Meta-Server	1smtA, winged helix (DNA binding) in domain 1
VNG5120H/6117H	-	Rosetta	small protein, 2 hits to 1asu00 phage integrase (weak hits)
VNG5122H/6119H	ISH7	Pfam	PF01609
VNG5123H/6120H	ISH7	TMHMM	membrane protein, unknown function
VNG5124H/6121H	-	-	-
IS-element rich region 2			
VNG5040H	ISH8	Pfam	PF01609
VNG5041H/5256H	-	Rosetta	-
VNG5042H/5255H Domain 1	ISH9	Rosetta	Hit to 1ez3A0, 2 long helices, no function annotation (domain 1)
VNG5042H/5255H Domain 2	ISH9	Pfam	PF01609 (domain 2)
VNG5044H/5253H	ISH5	Pfam	PF01609
VNG5045H/5252H	ISH11 (in ISH5)	Pfam	PF01609
VNG5047H/5250H	-	Rosetta	Hit to 1am3 (HIV capsid protein), 1.10.1200.30
VNG5048H/5249H	-	Rosetta	Hit to 1ais (cyclin-like fold/TBP fragment), 1.10.472.10
VNG5049H/5248H	-	Rosetta	Hit to 2ezh (transposase/transcription factor), 1.10.10.60
VNG5050H/5247H	-	Pfam	PF03551, PadR repressor
tbpB	-	known	tata-box binding protein B

encoded by the genes of this operon remain unresolved beyond the general function prediction afforded by the Rosetta and Pfam matches described above.

In the third case, the Rosetta predicted fold match to diphtheria toxin repressor (1bi2-B) for VNG0039H, as well as a weak hit by the FFAS method to the transcriptional regulator 1mkmA, suggest a winged helix fold and a transcriptional regulatory function or DNA binding function. VNG0039 has no edges in the association network and is not a member or a predicted operon. Again, further interpretation of the specific role of VNG0039H in the cell is dependent on combining this structure prediction with other systems-wide experiments aimed at inferring the global genetic regulatory circuit.

Ultimately we will have to test all proposed regulators, detected by sequence or structure, by performing gene knockouts followed by microarray analysis or by directly localizing the binding sites of these proteins within the genome [68]. Thus, the predictions described in Table 3 represent a short list of possible regulators, and serve to narrow the number of proteins for which such costly experiments will have to be performed.

Conclusion

We have presented a genome-wide analysis that integrates data from a wide variety of methods including *de novo* protein structure prediction, to provide unprecedented coverage

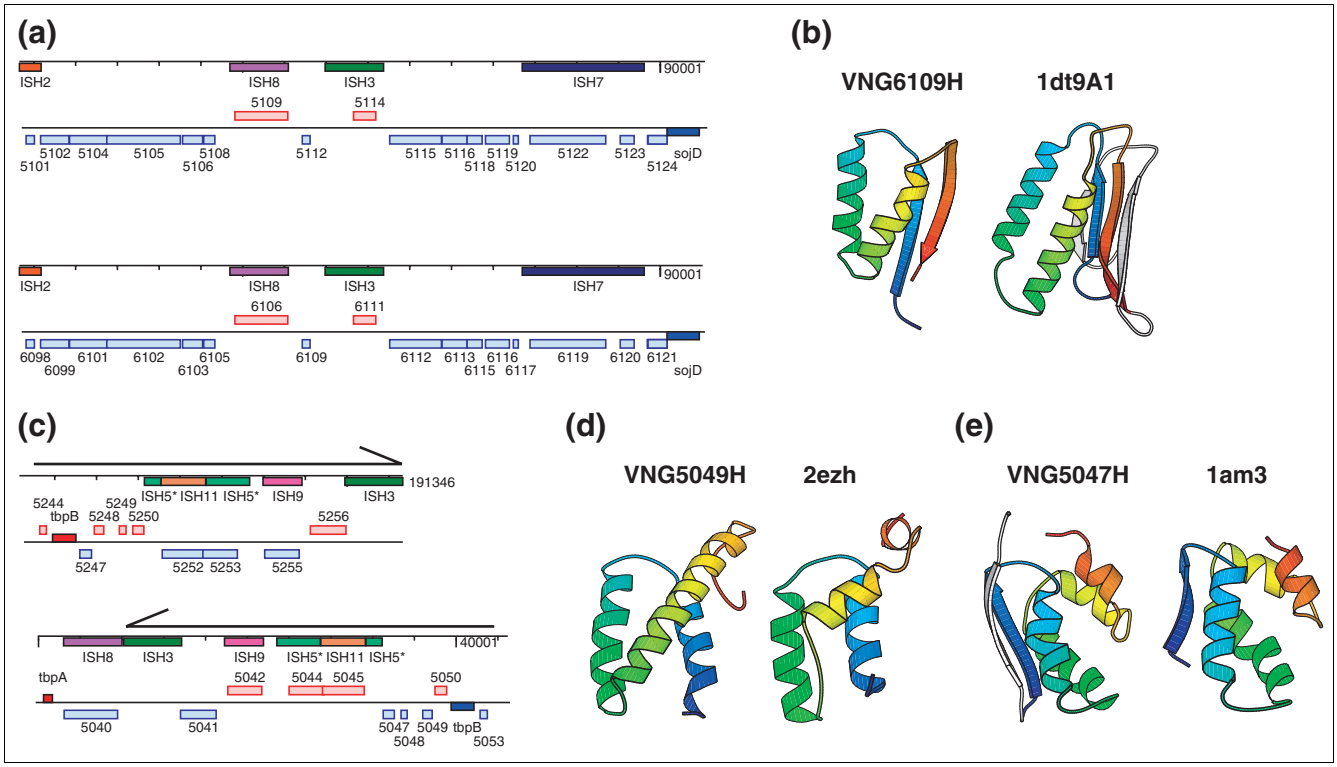

Figure 3

IS-element (insertion sequences) rich regions on the minichromosome. **(a)** Segment of *Halobacterium* genome corresponding to genes VNG5101H - sojD (and duplicate region VNG6098H - sojD). IS-elements are shown above as colored boxes. Open reading frames are indicated as red/pink (on 3' strand) or blue/sky-blue boxes (on 5' strand). **(b)** Top ranked Rosetta prediction for VNG6109H shown next to its closest match in the PDB, 1dt9A1 (translation initiation factor sub-domain). **(c)** Segment of *Halobacterium* genome corresponding to genes VNG5244H - VNG5256H (duplicated on the opposite strand elsewhere on the minichromosome, VNG5053H - VNG5041H). **(d)** Top ranked Rosetta prediction for VNG5049H shown next to its closest hit in the PDB, 2ezh. **(e)** Top ranked Rosetta prediction for VNG5047H shown next to its closest hit in the PDB, 1am3.

and comprehension of an important model archaeal system that is central to an ongoing multi-institute systems biology effort. We have shown that sources of putative annotation, such as Rosetta *de novo* fold predictions and phylogenetic profile predictions, become vastly more powerful when integrated with the full repertoire of applicable methods and presented to biologists in a well-collated and navigable environment. We would also like to note here that this database of sequence and structure-based annotations was used, in combination with the association network, to interpret the results of a recent study of the global response of *Halobacterium NRC-1* to ultraviolet radiation (providing insights into proteins of unknown function that were part of the global response to ultraviolet radiation) [3]. Central to the acceptance of this work by the biological community is the flexible data integration and visualization capabilities afforded by Cytoscape and SBEAMS. Although we have shown here, and elsewhere, examples of how this annotation system can generate testable biological hypotheses, an additional benefit of this work will be realized by making the results, in the form of the fully integrated dataset, publicly available. We also plan to perform this analysis for a broader array of genomes (yeast, selected human and mouse proteins, *Haloarcula marismortui*, etc) and thus further expand the utility of this approach.

Due to the errors inherent in current structure prediction and domain parsing methods there are still many misannotations, missed domain boundaries, and proteins with no annotation in the *Halobacterium NRC-1* genome, illustrating the need for improvements in structure prediction methods and domain parsing methods. Additionally, the use of Rosetta requires large amounts of computer processing time compared to sequence-based and fold-recognition methods (although Moore's law is rapidly making the computational cost of *de novo* prediction a less pressing issue). In spite of these caveats, integrating protein structure prediction with several orthogonal sources of general and putative functional information has allowed us to generate an experimentally testable hypothesis for significant numbers of proteins of otherwise unknown function.

Table 3

Predicted transcriptional regulators

Protein	Cluster number	CATH ID	Z-score	Confidence	Other hits
Winged helix repressor DNA binding domain					
VNG0389C	1	1.10.10.10	6.55	0.343	Weak FFAS hit to 1mzbA (ferric uptake gene regulator)
	2	1.10.10.10	7.40	0.4	
	4	1.10.10.10	6.76	0.357	
VNG2614H	1	1.10.10.10	6.64	0.313	PSI-BLAST to 1jgsA (winged helix, MarR)
	4	1.10.10.60	7.12	0.419	
	5	1.10.10.10	8.09	0.469	
VNG0768H	1	1.10.10.10	7.14	0.385	Not applicable
	9	1.10.10.10	8.09	0.481	
VNG2369C	1	1.10.10.10	7.14	0.323	Not applicable
	5	1.10.10.10	7.69	0.424	
VNG2641H	1	1.10.10.10	7.58	0.321	Not applicable
VNG1640H	1	1.10.10.10	7.86	0.34	Not applicable
VNG5156H	2	1.10.10.10	7.94	0.401	Weak FFAS hit to 1i1gA (LRP-like transcriptional regulator)
	3	1.10.10.10	9.36	0.502	
	4	1.10.10.10	8.55	0.444	
VNG5108H	14	1.10.10.10	8.04	0.388	Not applicable
VNG6047H	1	1.10.10.10	8.15	0.446	Weak FFAS hit to 1id3D (histone fold)
	3	1.10.10.60	8.69	0.524	
	14	1.10.10.10	10.03	0.58	
VNG0703H	1	1.10.10.10	8.17	0.439	PSI-BLAST to 1i1gA (LPR-like regulator)
	2	1.10.10.10	7.56	0.37	
	4	1.10.10.10	7.88	0.428	
	5	1.10.10.10	7.16	0.369	
	6	1.10.10.10	8.96	0.505	
VNG0462C	1	1.10.10.10	8.42	0.527	PSI-BLAST to 1lnwA (mexR repressor, winged helix fold) PSI-BLAST to ArsR
	2	1.10.10.10	9.77	0.621	
	4	1.10.10.10	7.66	0.489	
	5	1.10.10.10	8.57	0.545	
VNG2014H	14	1.10.10.10	8.72	0.379	Not applicable
VNG1229H	1	1.10.10.10	8.80	0.463	Not applicable
	4	1.10.10.10	7.95	0.339	
VNG0837H	2	1.10.10.10	9.02	0.476	Not applicable
	4	1.10.10.10	7.53	0.377	
	5	1.10.10.10	8.39	0.436	
VNG6479H	1	1.10.10.10	9.29	0.53	Not applicable
VNG0039H	1	1.10.10.10	9.36	0.478	
	2	1.10.10.10	8.60	0.438	Weak FFAS hit to 1mkmA (transcriptional regulator IclR, amino-terminal domain)
	3	1.10.10.10	9.28	0.438	
	5	1.10.10.10	9.96	0.522	
Homeodomain-like superfamily					
VNG0293H	3	1.10.10.60	6.14	0.395	PSI-BLAST to 1i1gA (LPR-like regulator)
	4	1.10.10.10	7.01	0.427	
	5	1.10.10.10	6.63	0.393	
VNG2074H	11	1.10.10.60	7.09	0.354	Not applicable
	1	1.10.10.60	6.03	0.286	

Table 3 *(Continued)*

Predicted transcriptional regulators					
VNG6251H	3	1.10.10.60	8.06	0.378	TMHMM predicts 1 TM helix
Cyclin A, domain 1 fold					
VNG2133H	2	1.10.472.10	8.25	0.351	Not applicable
VNG6287H	15	1.10.472.10	9.14	0.306	Weak FFAS hit to 1smt-A (SMTB repressor)
	4	1.10.472.10	8.93	0.283	
	2	1.10.472.10	8.50	0.269	
VNG0511H	8	1.10.472.10	9.25	0.512	PSI-BLAST hit to 1i1gA
VNG1865H	8	1.10.472.10	9.55	0.365	Not applicable
	4	1.10.472.10	7.18	0.18	

Material and methods

Protein selection and domain parsing

Rosetta predictions were generated for proteins and protein domains lacking hits to known structure after parsing proteins into domains using Ginzu [37]. Ginzu implements a hierarchically organized combination of sequence based methods (primarily PSI-BLAST and Pfam) to separate proteins into domains and to reannotate regions of sequence based on sequence homologies not present/detectable when the genome was initially sequenced [69,70]. Ginzu first uses PSI-BLAST to search for hits between the PDB and the query sequence. If hits are found, regions of the query sequence corresponding to the PDB hit(s) are masked and the remaining regions are searched against Pfam using HMMER. Hits to Pfam are then masked and remaining regions are searched against NCBI's 'nr' sequence set using PSI-BLAST. Multiple sequence alignments resulting from this final PSI-BLAST run (if homologs to the nr database are found) are then parsed into domains (if possible) using the chili-eye-ball algorithm [37]. Ginzu results primarily in a domain parsing of the query sequence but also results in sequence homology based hits to the PDB and domain annotations via Pfam. TMHMM [71] was also run on the *Halobacterium* genome to detect transmembrane regions. All of these sequence-based methods were run in-house with the exception of TMHMM.

Rosetta structure prediction and structure-structure searches

For each query sequence 9,000 independent simulations were carried out, each one resulting in a unique low energy conformation. This ensemble of conformations was then clustered and ranked as previously described [29]; this resulted in 20 models for each query. These 20 models were then searched against the PDB using mammoth. For each model, mammoth produces a top-ranked match to the PDB and a Z-score for that match. The mammoth Z-score is related to the closeness of the match between the predicted structure and the experimental structure and the length of that structural alignment [14].

Two confidence functions to predict the success of Rosetta predictions were used in this study. The training of these confidence functions on an extensive benchmark is described previously [29]. To predict success, defined as correct structural superfamily identification, we used a confidence function which takes as inputs the Z-score of the best structure-structure match [14] to the PDB for a given prediction as well as the length and the simulation convergence as predictors of success.

The probability that one of our top five automatically generated fold predictions for a sequence is correct is a function of the best Z-score to the PDB for the top five cluster centers (Z), the simulation convergence or cluster threshold (C), and the protein length (L) as follows:

$$\log\left(\frac{p}{1-p}\right) = 0.527(Z) + 0.012(L) - 0.239(C) - 4.97$$

Additionally, we can estimate the probability (p) that an individual fold-match between a single model and its closest match in the PDB is correct. To do this we use the Z-score of the individual model's best match to the PDB (Z), the degree of simulation convergence (C), the length of the query protein (L), and the ratio of the lengths of the hit in the PDB to the length of the query (L_H/L_Q) as follows:

$$\log\left(\frac{p}{1-p}\right) = 0.416(Z) + 0.00982(L) - 0.326(C) - 1.01(L_H/L_Q) - 2.17$$

Fold recognition

Key domains of unknown function (domains longer than 150 residues) were also submitted to the Meta-Server/3D-jury [72-74]. Proteins were submitted to the MetaServer if too large for analysis by Rosetta, or when Rosetta produced low confidence or otherwise useless results for a given protein. Protein regions with homologies to known structure were modeled using Modeler and ModBase [75]. Several proteins were also analyzed using FFAS03 [76,77].

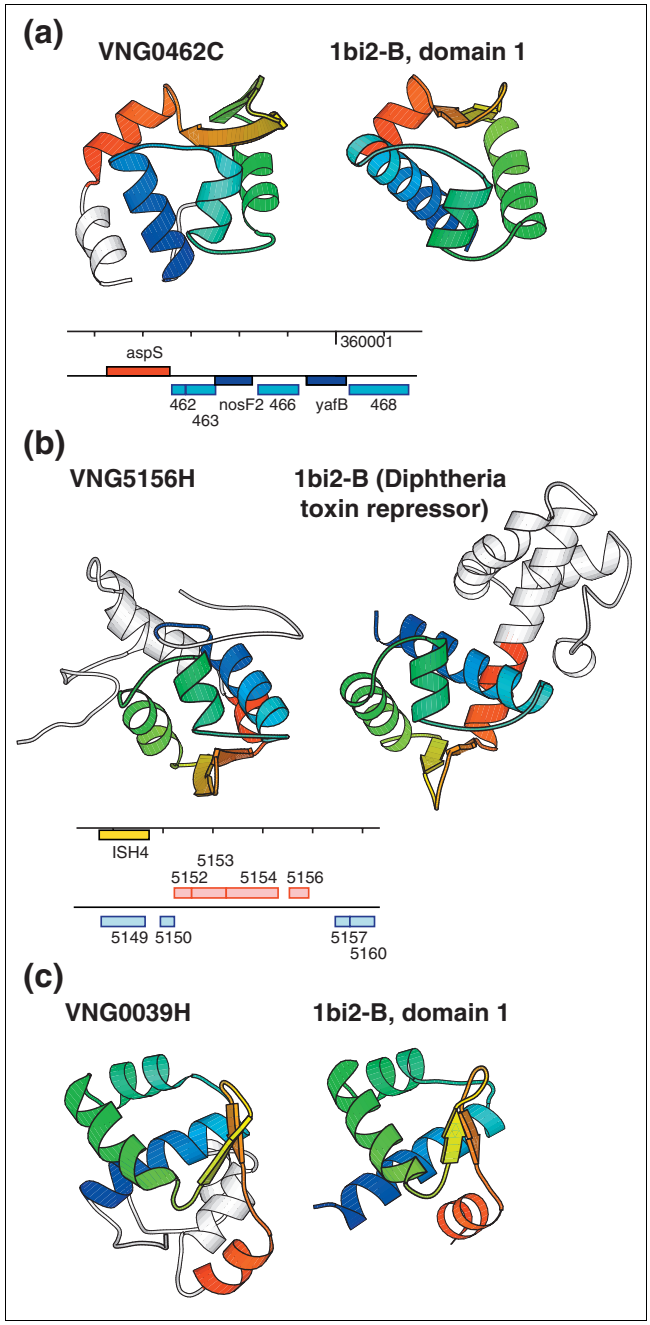

Figure 4
Predicted transcriptional regulators. Rosetta predictions for three *Halobacterium NRC-1* proteins that are consistent with transcription regulation and/or DNA binding. **(a)** The top ranked Rosetta structure prediction for VNG0462C (according to the Rosetta confidence function) is shown next to the diphtheria toxin repressor (1bi2-B). The predicted operon for VNG0462 is shown below; red/pink boxes above the line in this diagram are genes on the 3' strand while genes indicated by rectangles below the line are genes on the 5' strand. **(b)** The top ranked model for VNG5156H (left) is shown next to 1bi2-B, the predicted operon containing VNG5156, VNG5154, VNG5153 and VNG5152 is shown below. **(c)** The top ranked Rosetta prediction for VNG0039H is shown next to its closest match in the PDB, 1bi2-B.

Biological network construction

Operon prediction and co-regulated group edges

Edge types aimed at uncovering co-regulation patterns were calculated based on chromosomal proximity; we used two simple methods to predict conserved or significant chromosomal proximity. These methods were aimed at uncovering statistically significant chromosomal proximity and not the prediction of operons. One edge type indicates conserved patterns of proximity. Two proteins were given a chromosomal-proximity link if they were within 300 bp of each other in *Halobacterium* and had orthologs in at least one other organism within 300 bp [47]. This method has the disadvantage that it cannot give us insight into the function of proteins without orthologs in other systems. For proteins not grouped by the above method we used the method of Moreno-Hagelsieb and Collado-Vides to predict significant proximities on the basis of distance alone, as previously described [48]. A total of 1,662 links were generated by these two operon prediction methods.

Domain fusion and phylogenetic profile edges

Domain fusion

Separate protein sequences in the *Halobacterium* genome that are found fused in a single ORF in other genomes are predicted to interact functionally and physically [44].

Phylogenetic pattern

For a given protein, the presence or absence of an ortholog in fully sequenced genomes is termed its phylogenetic profile/pattern. Two proteins with identical phylogenetic profiles are often functionally related [45]. Domain fusion and phylogenetic profile edges were taken from the Predictome database [78]. A total of 2,946 domain fusion edges and 2,169 phylogenetic profile edges were added to the Halobacterium network.

Protein-protein interaction mapping

Yeast two-hybrid and *H. pylori* interactions were mapped onto *Halobacterium* by orthology using the COG database. If a pair of *Halobacterium* proteins belong to COGs known to interact in the yeast two-hybrid data they are given a COG-inferred yeast two-hybrid edge. Predictions also measured in *H. pylori* (via a comprehensive target-bait experiment) were also mapped onto *Halobacterium* protein pairs using this method [42,43,79]. A total of 1,143 putative protein-protein interaction edges were mapped onto *Halobacterium* in this way. Comparative modeling of *Halobacterium* structures was also used to predict protein-protein interactions. SCOP (Structural Classification Of Proteins) classification for *Halobacterium* sp. proteins were identified, when possible, based on homology modeling [75,80]. The SCOP classification is then used to determine probable interactions based on the prior observation that some pairs of SCOP families are known to physically interact in protein complexes in the PDB with a disproportionately high frequency [81]. The SCOP classifica-

tion of *Halobacterium* sp. proteins has been used to identify 562 likely SCOP-based interactions.

Data visualization

Cytoscape is a network visualization and exploration tool that allows for the simultaneous visualization of microarray, proteomics and network data. Cytoscape also has links to several external sources of annotation (KEGG, GO, etc.) and a broad array of integrated programs (Biomodule calculation, simulation of regulatory dynamics, network aware promoter and protein motif recognition, etc.) [49]. Cytoscape is currently a open-source project being developed at several institutions. See [82] for more information and for access to executables and source-code. SBEAMS was used as the database wherein the multitude of data-types described in this work were stored, maintained and accessed throughout the annotation process. SBEAMS provides convenient facilities for HTML display of tables it contains, with seamless access to external databases through a web interface. SBEAMS will be our primary portal through which data are released to the public. Additionally, external users accessing the database can add comments to any annotation/row in the table, thus allowing us to reconcile/correct our annotation after initial release.

Acknowledgements

We acknowledge the support of NSF-0220153 and NSF-0313754 to NSB and LH and DOD DAAD-13-03-C0057 to PS. We thank Dylan Chivian, David Kim and David Baker for use of the Ginzu program. We thank Erik Schweighofer, Ron Pankiewicz, Kerry Deutsch and Andrew Peabody for their help meeting the high performance computing needs of this work. We thank Jared Roach for discussion and comments related to this manuscript.

References

1. DasSarma S, Fleischmann EM: *Halophiles* Plainview, NY: Cold Spring Harbor Laboratory Press; 1995.
2. McCready S, Marcello L: **Repair of UV damage in *Halobacterium salinarum*.** *Biochem Soc Trans* 2003, **31:**694-698.
3. Baliga NS, Bjork SJ, Bonneau R, Pan M, Iloanusi C, Kottemann MC, Hood L, DiRuggiero J: **Systems level insights into the stress response to UV radiation in the halophilic archaeon *Halobacterium NRC-1*.** *Genome Res* 2004, **14:**1025-1035.
4. Rost B, Valencia A: **Pitfalls of protein sequence analysis.** *Curr Opin Biotechnol* 1996, **7:**457-461.
5. Devos D, Valencia A: **Practical limits of function prediction.** *Proteins* 2000, **41:**98-107.
6. Devos D, Valencia A: **Intrinsic errors in genome annotation.** *Trends Genet* 2001, **17:**429-431.
7. Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, *et al.*: **Genome sequence of *Halobacterium* species *NRC-1*.** *Proc Natl Acad Sci USA* 2000, **97:**12176-12181.
8. Fetrow JS, Skolnick J: **Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases.** *J Mol Biol* 1998, **281:**949-968.
9. Orengo CA, Todd AE, Thornton JM: **From protein structure to function.** *Curr Opin Struct Biol* 1999, **9:**374-382.
10. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective.** *J Mol Biol* 2001, **307:**1113-1143.
11. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C: **SCOP: a structural classification of proteins database.** *Nucleic Acids Res* 2000, **28:**257-259.
12. Holm L, Sander C: **Dali/FSSP classification of three-dimensional protein folds.** *Nucleic Acids Res* 1997, **25:**231-234.
13. Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM: **Protein folds and functions.** *Structure* 1998, **6:**875-884.
14. Ortiz AR, Strauss CE, Olmea O: **MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.** *Protein Sci* 2002, **11:**2606-2621.
15. Holm L, Sander C: **Protein structure comparison by alignment of distance matrices.** *J Mol Biol* 1993, **233:**123-138.
16. Orengo CA, Pearl FM, Thornton JM: **The CATH domain structure database.** *Methods Biochem Anal* 2003, **44:**249-271.
17. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247:**536-540.
18. Kinch LN, Grishin NV: **Evolution of protein structures and functions.** *Curr Opin Struct Biol* 2002, **12:**400-408.
19. Grishin NV: **Fold change in evolution of protein structures.** *J Struct Biol* 2001, **134:**167-185.
20. Rost B: **Protein structures sustain evolutionary drift.** *Fold Des* 1997, **2:**S19-S24.
21. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D: **Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins.** *Proteins* 1999, **34:**82-95.
22. Bonneau R, Baker D: ***Ab initio* protein structure prediction: progress and prospects.** *Annu Rev Biophys Biomol Struct* 2001, **30:**173-189.
23. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE, Baker D: **Rosetta in CASP4: progress in *ab initio* protein structure prediction.** *Proteins* 2001:119-126.
24. Aloy P, Stark A, Hadley C, Russell RB: **Predictions without templates: new folds, secondary structure, and contacts in CASP5.** *Proteins* 2003, **53(Suppl 6):**436-456.
25. Fischer D, Rychlewski L, Dunbrack RL Jr, Ortiz AR, Elofsson A: **CAFASP3: the third critical assessment of fully automated structure prediction methods.** *Proteins* 2003, **53(Suppl 6):**503-516.
26. Bradley P, Chivian D, Meiler J, Misura KM, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, *et al.*: **Rosetta predictions in CASP5: successes, failures, and prospects for complete automation.** *Proteins* 2003, **53(Suppl 6):**457-468.
27. Kinch LN, Wrabl JO, Krishna SS, Majumdar I, Sadreyev RI, Qi Y, Pei J, Cheng H, Grishin NV: **CASP5 assessment of fold recognition target predictions.** *Proteins* 2003, **53(Suppl 6):**395-409.
28. Bonneau R, Tsai J, Ruczinski I, Baker D: **Functional inferences from blind *ab initio* protein structure predictions.** *J Struct Biol* 2001, **134:**186-190.
29. Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D: ***De novo* prediction of three-dimensional structures for major protein families.** *J Mol Biol* 2002, **322:**65-78.
30. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL: **Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins.** *Nucleic Acids Res* 1999, **27:**260-262.
31. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, *et al.*: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32 Database issue:**D138-D141.
32. Baliga NS, Pan M, Goo YA, Yi EC, Goodlett DR, Dimitrov K, Shannon P, Aebersold R, Ng WV, Hood L: **Coordinate regulation of energy transduction modules in *Halobacterium* sp. analyzed by a global systems approach.** *Proc Natl Acad Sci USA* 2002, **99:**14913-14918.
33. Goo YA, Yi EC, Baliga NS, Tao WA, Pan M, Aebersold R, Goodlett DR, Hood L, Ng WV: **Proteomic analysis of an extreme halophilic archaeon, *Halobacterium* sp. *NRC-1*.** *Mol Cell Proteomics* 2003, **2:**506-524.
34. Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S: **Understanding the adaptation of *Halobacterium* species *NRC-1* to its extreme environment through computational analysis of its genome sequence.** *Genome Res* 2001, **11:**1641-1650.
35. Mallick P, Boutz DR, Eisenberg D, Yeates TO: **Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds.** *Proc Natl Acad Sci USA* 2002, **99:**9679-9684.
36. Thompson MJ, Eisenberg D: **Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability.** *J Mol Biol* 1999, **290:**595-604.
37. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D: **Automated prediction**

of CASP-5 structures using the Robetta server. *Proteins* 2003, **53(Suppl 6):**524-533.
38. **Halo Research at ISB** [http://halo.systemsbiology.net/]
39. **NCBI COGs** [http://www.ncbi.nlm.nih.gov/COG/]
40. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29:**22-28.
41. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, *et al.*: **The protein-protein interaction map of *Helicobacter pylori*.** *Nature* 2001, **409:**211-215.
42. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci USA* 2001, **98:**4569-4574.
43. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al.*: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403:**623-627.
44. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285:**751-753.
45. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: **Protein function in the post-genomic era.** *Nature* 2000, **405:**823-826.
46. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402:**86-90.
47. Mellor JC, Yanai I, Clodfelter KH, Mintseris J, DeLisi C: **Predictome: a database of putative functional links between proteins.** *Nucleic Acids Res* 2002, **30:**306-309.
48. Moreno-Hagelsieb G, Collado-Vides J: **A powerful non-homology method for the prediction of operons in prokaryotes.** *Bioinformatics* 2002, **18(Suppl 1):**S329-S336.
49. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
50. Brooun A, Bell J, Freitas T, Larsen RW, Alam M: **An archaeal aerotaxis transducer combines subunit I core structures of eukaryotic cytochrome c oxidase and eubacterial methyl-accepting chemotaxis proteins.** *J Bacteriol* 1998, **180:**1642-1646.
51. Schimz A, Hildebrand E: **Chemosensory responses of *Halobacterium halobium*.** *J Bacteriol* 1979, **140:**749-753.
52. Appleman JA, Stewart V: **Mutational analysis of a conserved signal-transducing element: the HAMP linker of the *Escherichia coli* nitrate sensor NarX.** *J Bacteriol* 2003, **185:**89-97.
53. Appleman JA, Chen LL, Stewart V: **Probing conservation of HAMP linker structure and signal transduction mechanism through analysis of hybrid sensor kinases.** *J Bacteriol* 2003, **185:**4872-4882.
54. Aravind L, Ponting CP: **The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins.** *FEMS Microbiol Lett* 1999, **176:**111-116.
55. Kokoeva MV, Storch KF, Klein C, Oesterhelt D: **A novel mode of sensory transduction in archaea: binding protein-mediated chemotaxis towards osmoprotectants and amino acids.** *EMBO J* 2002, **21:**2312-2322.
56. Ventura M, Foley S, Bruttin A, Chennoufi SC, Canchaya C, Brussow H: **Transcription mapping as a tool in phage genomics: the case of the temperate *Streptococcus thermophilus* phage Sfi21.** *Virology* 2002, **296:**62-76.
57. Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H: **Prophage genomics.** *Microbiol Mol Biol Rev* 2003, **67:**238-276.
58. Desiere F, Pridmore RD, Brussow H: **Comparative genomics of the late gene cluster from *Lactobacillus* phages.** *Virology* 2000, **275:**294-305.
59. Brussow H, Desiere F: **Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages.** *Mol Microbiol* 2001, **39:**213-222.
60. Tang SL, Nuttall S, Ngui K, Fisher C, Lopez P, Dyall-Smith M: **HF2: a double-stranded DNA tailed haloarchaeal virus with a mosaic genome.** *Mol Microbiol* 2002, **44:**283-296.
61. Stedman KM, She Q, Phan H, Arnold HP, Holz I, Garrett RA, Zillig W: **Relationships between fuselloviruses infecting the extremely thermophilic archaeon Sulfolobus: SSV1 and SSV2.** *Res Microbiol* 2003, **154:**295-302.
62. Ng WV, Ciufo SA, Smith TM, Bumgarner RE, Baskin D, Faust J, Hall B, Loretz C, Seto J, Slagel J, *et al.*: **Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome?** *Genome Res* 1998, **8:**1131-1141.
63. Ng WL, Kothakota S, DasSarma S: **Structure of the gas vesicle plasmid in *Halobacterium halobium* inversion isomers, inverted repeats, and insertion sequences.** *J Bacteriol* 1991, **173:**3933.
64. Boltner D, MacMahon C, Pembroke JT, Strike P, Osborn AM: **R391: a conjugative integrating mosaic comprised of phage, plasmid, and transposon elements.** *J Bacteriol* 2002, **184:**5158-5169.
65. Osborn AM, Boltner D: **When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum.** *Plasmid* 2002, **48:**202-212.
66. Sato S, Nakada Y, Shiratsuchi A: **IS421, a new insertion sequence in *Escherichia coli*.** *FEBS Lett* 1989, **249:**21-26.
67. Lewin B: *Genes* 7th edition. New York: Oxford University Press; 2000.
68. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al.*: **Transcriptional regulatory networks in *Saccharomyces cerevisiae*.** *Science* 2002, **298:**799-804.
69. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
70. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2000, **28:**263-266.
71. Sonnhammer EL, von Heijne G, Krogh A: **A hidden Markov model for predicting transmembrane helices in protein sequences.** *Proc Int Conf Intell Syst Mol Biol* 1998, **6:**175-182.
72. Lundstrom J, Rychlewski L, Bujnicki J, Elofsson A: **Pcons: a neural-network-based consensus predictor that improves fold recognition.** *Protein Sci* 2001, **10:**2354-2362.
73. **Bioinfo.pl Meta Server Job List** [http://bioinfo.pl/meta/]
74. **Pcons** [http://www.sbc.su.se/~arne/pcons/]
75. Pieper U, Eswar N, Stuart AC, Ilyin VA, Sali A: **MODBASE, a database of annotated comparative protein structure models.** *Nucleic Acids Res* 2002, **30:**255-259.
76. Rychlewski L, Jaroszewski L, Li W, Godzik A: **Comparison of sequence profiles. Strategies for structural predictions using sequence information.** *Protein Sci* 2000, **9:**232-241.
77. Jaroszewski L, Rychlewski L, Godzik A: **Improving the quality of twilight-zone alignments.** *Protein Sci* 2000, **9:**1487-1496.
78. **Predictome guide/FAQ** [http://predictome.bu.edu]
79. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, *et al.*: **The protein-protein interaction map of *Helicobacter pylori*.** *Nature* 2001, **409:**211-215.
80. Murzin AG: **Structure classification-based assessment of CASP3 predictions for the fold recognition targets.** *Proteins* 1999:88-103.
81. Park J, Lappe M, Teichmann SA: **Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast.** *J Mol Biol* 2001, **307:**929-938.
82. **Cytoscape** [http://www.cytoscape.org]