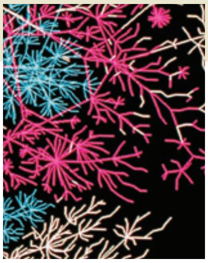


Detection of parallel functional modules by comparative analysis of genome sequences

Huiying Li[1], Matteo Pellegrini[1,2] & David Eisenberg[1]

Parallel functional modules are separate sets of proteins in an organism that catalyze the same or similar biochemical reactions but act on different substrates or use different cofactors. They originate by gene duplication during evolution. Parallel functional modules provide versatility and complexity to organisms, and increase cellular flexibility and robustness. We have developed a four-step approach for genome-wide discovery of parallel modules from protein functional linkages. From ten genomes, we identified 37 cellular systems that consist of parallel functional modules. This approach recovers known parallel complexes and pathways, and discovers new ones that conventional homology-based methods did not previously reveal, as illustrated by examples of peptide transporters in *Escherichia coli* and nitrogenases in *Rhodopseudomonas palustris*. The approach untangles intertwined functional linkages between parallel functional modules and expands our ability to decode protein functions from genome sequences.

Organisms maintain families of similar yet distinct gene sequences–so-called paralogs. Paralogs originated by gene duplication and evolved through a variety of gene-rearrangement mechanisms[1]. It has been shown that 50% of prokaryotic genes[2,3] and over 90% of eukaryotic genes[4] are generated from gene duplication. During evolution gene products may acquire new substrate or cofactor specificities, altered recognition properties, new interacting partners and otherwise modified functions[1]. Parallel functional modules may arise when a set of proteins that function together in a pathway or complex are duplicated and modified. The discovery of frequent gene duplication in genomes suggests that many genes encode proteins with at least partially redundant functions. This partial redundancy in the functions of parallel modules offers flexibility at the biochemical level, a buffering effect for the system and better adaptation of organisms to their environments.

Previously, Kelley *et al*[5]. identified paralogous pathways from protein interaction networks constructed from experimental data on budding yeast. They developed a network comparison method to graphically align protein networks from different organisms in a pair-wise fashion. To search for potential paralogous pathways, the authors aligned the yeast network to itself, and required that pairs of nodes (proteins) have sequence similarity with BLAST[6] E-values no greater than 10^{-10} and that the edges (interactions between proteins) are conserved between the paralogous pathways. Three hundred high-scoring alignments of paralogous pathways of length four were discovered in the study.

Here, we explore the possibility of detecting parallel functional modules directly from genome sequences. This approach can be applied to organisms whose genome sequences are available. To date, more than 200 genomes have been fully sequenced and several hundred genomes are in the process of being sequenced. Most of the genome sequences are readily available from public databases and represent a much broader spectrum of organisms than do large-scale experimental data, which are available only for a limited number of model organisms.

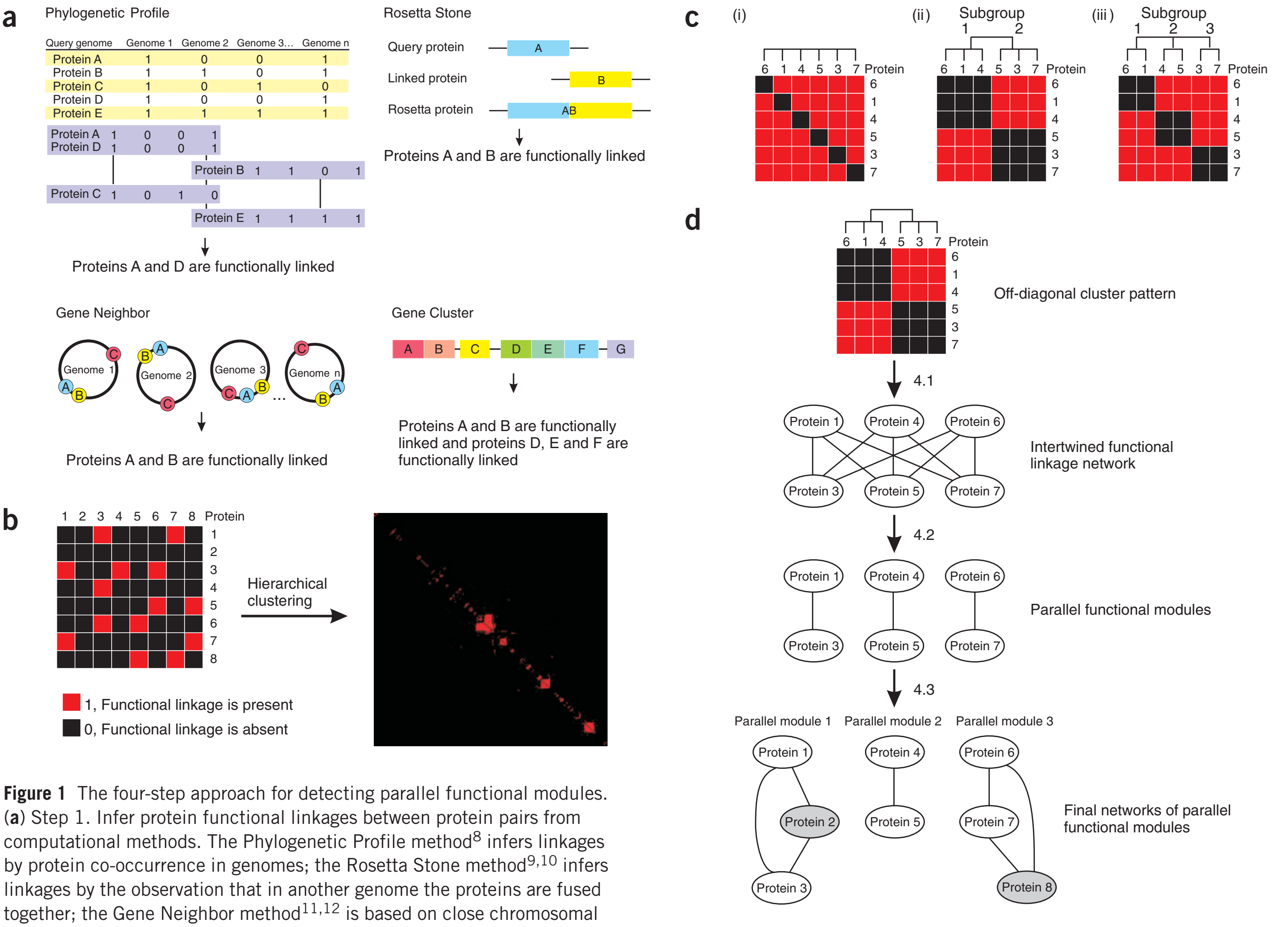
Several computational methods based on the genomic context have been developed to infer functionally linked proteins. It has been shown that the performance of these computational methods in inferring protein interactions is quantitatively comparable to, if not better than, the use of genomic-scale experimental data[7]. Among them are the Phylogenetic Profile method[8], which identifies the protein pairs that co-occur in various genomes, the Rosetta Stone method[9,10], which identifies the protein pairs that fuse into a single polypeptide in another organism, the Gene Neighbor method[11,12], which identifies the protein pairs that reside in close chromosomal proximity in multiple genomes, and the Gene Cluster method, which identifies the protein pairs that are likely to belong to the same operons based on the intergenic distances[13,14]. The availability of these computational methods makes it feasible to infer the proteins in a pathway or complex directly from genome sequences.

The four-step approach

Our approach for genome-wide discovery of parallel functional modules involves four steps (**Fig. 1** and **Supplementary Fig. 1** online). In step 1, we calculate functional linkages[15] for all possible protein pairs in the query genome by comparison with the proteins encoded in 82 other fully sequenced genomes[14] (**Supplementary Methods** online). Linkages are between the proteins from the same organism only. Functional linkages are calculated by the Phylogenetic Profile[8], Rosetta Stone[9,10], Gene Neighbor[11,12] and Gene Cluster methods[13,14] (**Fig. 1a**). The output of this step is a binary description of the linkage between every pair of proteins encoded in the query genome. If the two proteins are linked with a confidence above the chosen threshold, the linkage is 1; otherwise, the linkage is 0.

[1]Howard Hughes Medical Institute, UCLA-DOE Institute for Genomics and Proteomics, Department of Chemistry and Biochemistry, University of California, Los Angeles, Box 951570, Los Angeles, California 90095-1570, USA. [2]Present address: Rosetta Inpharmatics LLC, a wholly owned subsidiary of Merck & Co., Inc., 401 Terry Ave. N, Seattle, Washington 98109, USA. Correspondence should be addressed to D.E. (david@mbi.ucla.edu).

Published online 4 February 2005; doi:10.1038/nbt1065



Query genome	Genome 1	Genome 2	Genome 3...	Genome n
Protein A	1	0	0	1
Protein B	1	1	0	1
Protein C	1	0	1	0
Protein D	1	0	0	1
Protein E	1	1	1	1

Figure 1 The four-step approach for detecting parallel functional modules. (**a**) Step 1. Infer protein functional linkages between protein pairs from computational methods. The Phylogenetic Profile method[8] infers linkages by protein co-occurrence in genomes; the Rosetta Stone method[9,10] infers linkages by the observation that in another genome the proteins are fused together; the Gene Neighbor method[11,12] is based on close chromosomal positions of two gene-encoded proteins in various genomes; and the Gene Cluster method[13,14] infers linkages based on short intergenic distances between two genes in query genomes. (**b**) Step 2. Construct a symmetric matrix (number of genes × number of genes) to represent the inferred functional linkages between protein pairs (left). The proteins are then hierarchically clustered based on the pattern of their linkages to other proteins in the genome, generating a clustered genome-wide functional linkage map. The map shown on the right is a part of the *E. coli* K12 genome functional linkage map. The clustered map is composed of clusters with linked proteins arrayed near the diagonal (upper left to lower right). The chromosomal order on the rows and columns is lost after clustering. (**c**) Step 3. Search visually for off-diagonal cluster patterns—the signature of parallel functional modules—in the clustered functional linkage map. (**i**) A typical cluster pattern for pathways and complexes. (**ii**) An off-diagonal cluster pattern for three parallel functional modules each with two major components. (**iii**) An off-diagonal cluster pattern for two parallel functional modules each with three major components. Note that in (**ii**) and (**iii**) proteins within a subgroup are linked to proteins in the other subgroup(s), but not to each other. (**d**) Step 4. Manually extract the proteins and their functional linkages encoded in the off-diagonal cluster pattern from the map (step 4.1). Match module partners and remove linkages between parallel functional modules that arise from paralogous relationships using gene location relationships (for prokaryotic genomes) or coevolution relationships (for eukaryotic genomes) (step 4.2). Finally, proteins that are linked to module components but not included in the off-diagonal cluster are added (proteins 2 and 8 in shaded circles), yielding a functional linkage network for the parallel functional modules (step 4.3).

In step 2, we construct a matrix of functional linkages for the query genome and group proteins based on the similarity of their functional linkage patterns using a hierarchical clustering algorithm[16] (**Fig. 1b** and **Supplementary Methods** online). The symmetric matrix of functional linkages is composed of 0s and 1s, which are calculated in step 1, representing the absence or presence of functional linkages between protein pairs. Proteins are initially ordered sequentially in rows and columns according to their gene order on the chromosome from the first to the last gene. We remove the rows and columns that have no functional linkages to any other proteins in the genome to reduce the size of the matrix. Next, we hierarchically cluster the matrix columns and rows based on the correlation coefficients[16]. The output is a new matrix with reordered columns and rows where proteins with the same cellular functions or in the same pathways or complexes cluster together. The reordered matrix can be visualized as a heat map, which we named a 'clustered genome-wide functional linkage map'[17]. The map is composed mostly of small clusters of highly linked proteins (**Fig. 1b**). Typically, clusters show up on the diagonal of the map (arbitrarily from upper left to lower right in our map) because of the symmetry of the functional linkage matrix (**Fig. 1c** (**i**)).

In step 3, we visually search for off-diagonal clusters within the functional linkage map. The off-diagonal cluster pattern is a signature of parallel functional modules, as explained below (**Fig. 1c** (**ii**) and (**iii**)). Unlike typical clusters, which consist of one group of proteins with similar functional linkages, off-diagonal clusters consist of two or more distinct subgroups of proteins. Each subgroup is a collection of the equivalent, usually paralogous, components in parallel functional modules. Proteins in the same subgroup usually are not functionally linked to each other

Table 1 Thirty-seven cellular systems of parallel functional modules from ten genomes

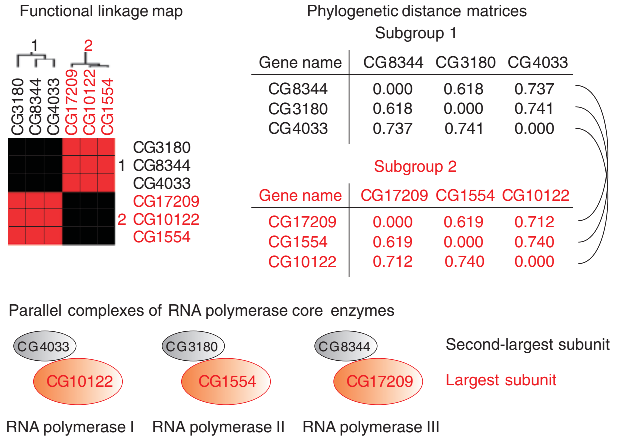
Genomes		Protein clusters with parallel functional modules	Previously described
Bacteria	*Escherichia coli* K12	1. Peptide transporters[1]	1. Partially described[28–32]
		2. Ribonucleoside diphosphate reductases	
		3. Heat shock proteins	3. Described[41]
		4. Pyruvate dehydrogenases	
	Escherichia coli O157:H7	1. Tail assembly proteins	
		2. Tail assembly chaperones	
		3. Sugar transporters	
		4. Branched-chain amino acid transporters	
	Rhodopseudomonas palustris	1. Nitrogenases[1]	1. Partially described[36]
		2. Aromatic compound degradation pathways	2. Described[36]
		3. Multidrug resistance efflux pumps	
		4. Transposases	
		5. Glutamate synthases	
		6. Two-component systems	6. Described[36]
	Sinorhizobium meliloti	1. Transposases	1. Described[42]
		2. DNA replication proteins	
		3. Flagellar proteins	
		4. hemK	
Archaea	*Pyrobaculum aerophilum*	1. Ferredoxin oxidoreductases	
		2. Carbon monoxide dehydrogenases	
	Methanocaldococcus jannaschii	1. Restriction enzymes	1. Described[43], not all found
Eukaryotes	*Saccharomyces cerevisiae*	1. Helicases	
		2. Heat shock proteins	2. Described[44]
		3. Unknown pathways	
	Caenorhabditis elegans	1. Myosins	1. Described[45]
		2. Pyridoxal phosphate-dependent enzymes	
		3. ATP synthetases	
		4. Monocarboxylate transporters	4. Partially described[46]
		5. Heat shock proteins	5. Described[47]
		6. Unknown pathways	
	Drosophila melanogaster	1. RNA polymerases[1]	1. Described[22–26]
		2. Rab3GEF	
		3. Heat shock proteins	3. Partially described[48,49]
	Arabidopsis thaliana	1. Pectinesterases	
		2. Glucanases	2. Described[50]
		3. Heat shock proteins	3. Described[51]
		4. Unknown pathways	

The modules were revealed by off-diagonal cluster patterns in the functional linkage maps. Each cellular system contains two or more parallel functional modules.

[1] Details of the RNA polymerases from *D. melanogaster*, peptide transporters from *E. coli* K12 and nitrogenases from *R. palustris* are illustrated in **Figures 2–5**.

as indicated by the black squares in **Figure 1c** (**ii**) and (**iii**). Proteins in different subgroups, however, are functionally linked as indicated by the off-diagonal red squares in **Figure 1c** (**ii**) and (**iii**). Functional linkages among the proteins from different subgroups suggest that the proteins in one subgroup are functional partners in a pathway or complex of the proteins in the other subgroup(s). As the methods used to infer functional linkages do not always distinguish between paralogs, a protein is often linked not only to its pathway or complex partner but also to the paralogs of that partner. This results in tangled linkages between parallel functional modules. Many of these linkages are spurious and need to be removed.

In step 4, to untangle the intertwined functional linkages among parallel functional modules, we manually match the module partners from each subgroup and remove functional linkages that result from paralogous relationships (**Fig. 1d**). In prokaryotic genomes, the products of genes that reside near each other on the chromosome are likely to interact with each other within a pathway or complex[11,12]. We thus pair the proteins whose genes are located in the same region on the chromosome, sometimes in the same operon, from different subgroups. All the proteins in a functional linkage map are from the same genome. Therefore, it can be determined whether two genes are in the same region of the chromosome, meaning that the genes are neighbors but do not necessarily belong to the same operon. We infer whether two genes are in the same operon by computing their intergenic distances and the probability of observing the genes belonging to the same operon[14]. For eukaryotic genomes, we compare the phylogenetic distance matrices of the protein sequences from the two subgroups, and pair the proteins with the closest phylogenetic distances[18,19] (**Supplementary Methods** online) as explained in detail below in the example of RNA polymerases in *Drosophila melanogaster*. The assumption underlying this distance-based approach is that



Subgroup 1

Gene name	CG8344	CG3180	CG4033
CG8344	0.000	0.618	0.737
CG3180	0.618	0.000	0.741
CG4033	0.737	0.741	0.000

Subgroup 2

Gene name	CG17209	CG1554	CG10122
CG17209	0.000	0.619	0.712
CG1554	0.619	0.000	0.740
CG10122	0.712	0.740	0.000

Figure 2 RNA polymerases display a parallel complex pattern in the clustered genome-wide functional linkage map of *D. melanogaster*. The map shown on the upper left is a symmetric matrix with proteins in the rows and columns in the same order. Each red square corresponds to the presence of a functional linkage between the protein in the row and the protein in the column, and each black square corresponds to the absence of a functional linkage. All the largest and second-largest subunits of RNA polymerases I, II and III are clustered together. The second-largest subunits form subgroup 1 and the largest subunits form subgroup 2. Within each subgroup, the three proteins are from RNA polymerase I, II and III, respectively. Without knowing the protein functions except the gene names, we generated the phylogenetic distance matrices for the proteins in the two subgroups shown on the upper right panel. The numbers in the matrices represent the sequence distances between the protein pairs. We aligned the two matrices to identify interacting pairs. The three resulting complexes turned out to be the correct complexes of RNA polymerases I, II and III core enzymes.

proteins that function together evolve at similar rates during evolution. By aligning the phylogenetic trees of two groups of interacting proteins, we can often identify the correct pairings between the group members. For the prokaryotic genomes whose operon information is not sufficient, the phylogenetic distance method can also be applied to pair the proteins. After we match functional module partners from each subgroup, we remove the links between parallel functional modules that arise from paralogous relationships. Thus, intertwined networks are separated into parallel functional modules. To make the modules complete, we fill in the proteins that are functionally linked to other proteins in the module but are not included in the off-diagonal cluster. The final product of this approach is a refined protein interaction network revealing parallel functional modules encoded in the genome (**Fig. 1d**).

Using the four-step approach, we analyzed the genomes of ten organisms, including four bacteria (*Escherichia coli* K12, *Escherichia coli* O157: H7, *Rhodopseudomonas palustris*, *Sinorhizobium meliloti*), two archaea (*Pyrobaculum aerophilum*, *Methanocaldococcus jannaschii*), and four eukaryotes (*Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*). The results are summarized in **Table 1**. We discovered 37 cellular systems that have two or more parallel functional modules, many of which are novel. Some of the parallel functional modules are found in multiple organisms. To our knowledge, ten of the 37 cellular systems have been described in the literature. Another four have been partially described previously but we found new members of them with our analysis. The functional modules in one previously described cellular system were partially recovered in our analysis. The rest (60%) of the 37 cellular systems with parallel functional modules that we identified have not been previously detected in these organisms, as judged by a PubMed search. Three examples of parallel functional modules are discussed below in detail.

RNA polymerases in *Drosophila melanogaster*

As an example to illustrate our four-step approach, we start with the well-characterized case of RNA polymerases. Biochemical and genetic analyses have established that eukaryotic RNA polymerases I, II and III are evolutionarily related[20]. The core enzymes of these polymerases are composed of two large subunits that are homologous to the two largest subunits of prokaryotic RNA polymerases[21]. The *D. melanogaster* genome functional linkage map revealed an off-diagonal cluster pattern composed of two subgroups of three genes (**Fig. 2**). One subgroup is composed of the second-largest subunits of RNA polymerases I, II, III[22,23] and the other subgroup contains the largest subunits of RNA polymerases[24–26]. To match the largest subunits with their cognate core enzyme partners, we calculated the phylogenetic distance matrices and aligned the two matrices to identify the interacting pairs (**Fig. 2**). The three pairs of RNA polymerase subunits are correctly matched. Therefore, in this example of our approach, we not only detect these parallel complexes in the genome of *D. melanogaster* but also correctly assign the components of each complex. A similar result was found by Kelley *et al.* using their network alignment–based method in yeast[5].

Peptide transporters in *Escherichia coli* K12

Three previously known peptide transporters and three unknown transporters were revealed as parallel complexes by an off-diagonal cluster pattern on the functional linkage map of *E. coli* K12 (**Fig. 3a**). These proteins belong to the 79-member ATP-binding cassette (ABC) protein family, the largest paralogous family of proteins in *E. coli*[27]. The three previously known peptide transporters are Opp[28], Dpp[29] and Sap[30,31] proteins. A nickel transporter (Nik proteins) is also included in the cluster because of its sequence similarity to the peptide transporters[32]. In addition, we found three sets of uncharacterized proteins in the cluster (B1483-B1487, B0829-B0832, YejABEF). These proteins had been annotated in the NCBI database (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html) as putative transport proteins without specified substrates. Because these complexes are parallel to the known peptide transporter complexes, and share a similar operon structure as well as a similar network of functional linkages (**Fig. 3b**), we infer that these proteins are putative peptide transporters. The seven transporters found in the off-diagonal cluster are complete members of ABC subfamily 2 in *E. coli*, classified based on specific transmembrane domains and periplasmic-binding proteins of the transporters[27]. Thus, the four-step approach was able to separate the seven ABC transporters out of 79 paralogs in the *E. coli* genome. The result is consistent with the classification of ABC transporters based on sequences. In this case, we can also infer the individual role of each protein in the previously uncharacterized transporter complexes based on their membership within the observed subgroups. The peptide transporter cluster shows three subgroups. The three subgroups correspond to three major components of the transporter complexes and match the predicted cellular localizations of those components (**Fig. 3a**). For example, the first subgroup includes all the B proteins of each transporter, and they are all predicted by PSORT[33] to reside in the inner membrane. These B proteins probably all function as the permease in the transporter complexes.

Nitrogenases in *Rhodopseudomonas palustris*

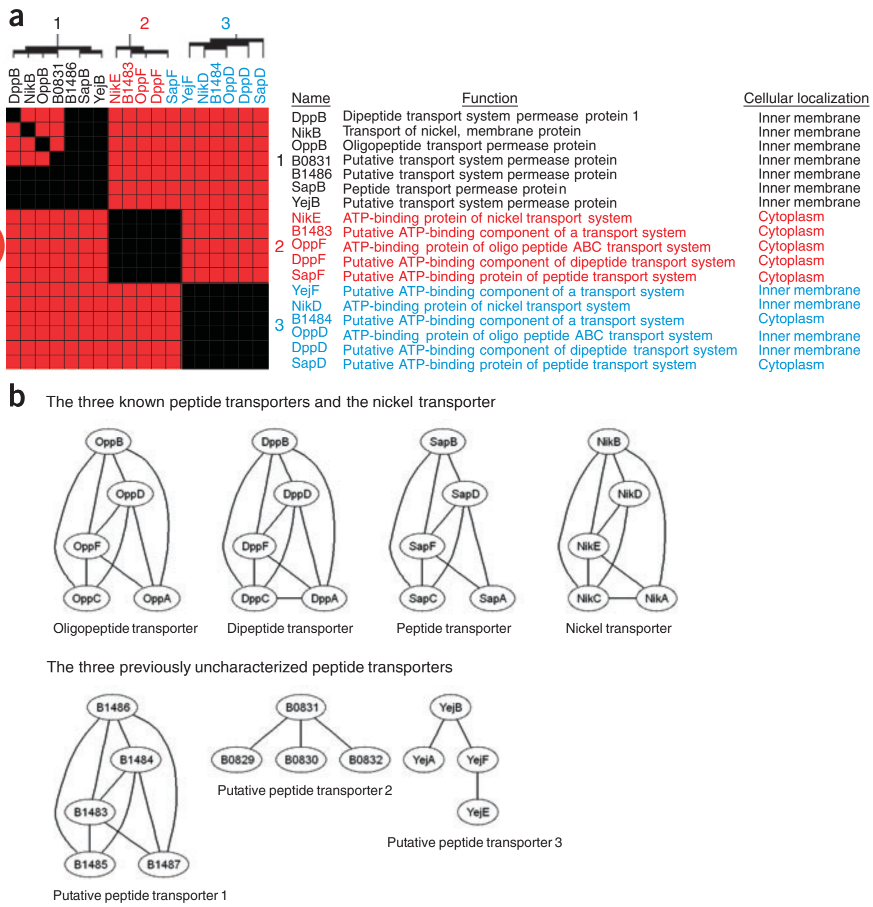
In *R. palustris*, one of the most metabolically versatile bacteria, we found two new putative nitrogenases in addition to three known ones (**Fig. 4a**). Previously only three types of nitrogenases have been found in all N_2-fixing organisms: Mo-Fe nitrogenase (*nif* genes), V-Fe nitrogenase (*vnf* genes) and Fe-Fe nitrogenase (*anf* genes), classified according to the cofactors in their active sites[34,35]. The two new sets of putative nitrogenases that we labeled as Xnf and Ynf were not turned up in a previously reported genome annotation of *R. palustris*[36]. Two neighboring *ynf* genes were annotated as homologs of the proteins from two different types of nitrogenases: the cofactor synthesis protein N of V-Fe nitrogenase and the α-subunit of Mo-Fe nitrogenase. In our analysis, these two genes are assigned as the α- and β-subunits of Ynf nitrogenase based on the parallel functional module pattern. This example demonstrates that pure homology-based methods can miss new members of parallel functional modules when the sequence similarity between members becomes less certain.

Sequence analysis of the new putative nitrogenase proteins suggests that these previously uncharacterized proteins are likely to form functional nitrogenases or nitrogenase-like protein complexes. The gene components of the putative new nitrogenases are similar to those of the three known nitrogenases (**Fig. 4b**). They all have the necessary structural genes for a functional nitrogenase: the α- and β-subunits of the dinitrogenase and the nitrogenase reductase (the iron protein). The cofactor synthesis proteins E share sequence and structure similarity with the α-subunits and their genes probably evolved from the same ancestor[37]. Based on the protein sequence alignment, we constructed a protein tree of the α-subunits and the cofactor synthesis proteins E of nitrogenases from *R. palustris* and *Azotobacter vinelandii*, which is shown in **Figure 4c**. The α-subunit of each known nitrogenase from *R. palustris* is clustered with the corresponding one from *A. vinelandii*. The α-subunits of the two new putative nitrogenases from *R. palustris* are clustered together as a separate branch from the other α-subunits. A key residue, Hisα442, is conserved in all three known nitrogenases[34]. It is a ligand to the FeMo cofactor in the active site of the Mo-Fe protein[34,35]. This residue is missing in the new putative nitrogenases; however, there are other histidine residues available around the equivalent sequence position of Hisα442. It is possible that the new putative nitrogenases may use different metals or cofactors for catalytic activity or they may perform some new biochemical activity evolved after gene duplication. These computational leads can be pursued experimentally to determine the specific functions and cofactors for these new putative nitrogenase proteins.

The five parallel sets of nitrogenase proteins are clearly depicted by the inferred functional linkage networks shown in **Figure 5**. Before untangling the functional linkages among parallel functional modules, one would have confronted a massive network of nitrogenase proteins containing many spurious linkages as shown in **a**. After applying the four-step approach, the network relationships among the nitrogenase proteins become clear, including their module partners and their paralogs in other modules. Notice that some cofactor synthesis proteins can be shared among different nitrogenases[38]. This could be true for the new nitrogenases as well.

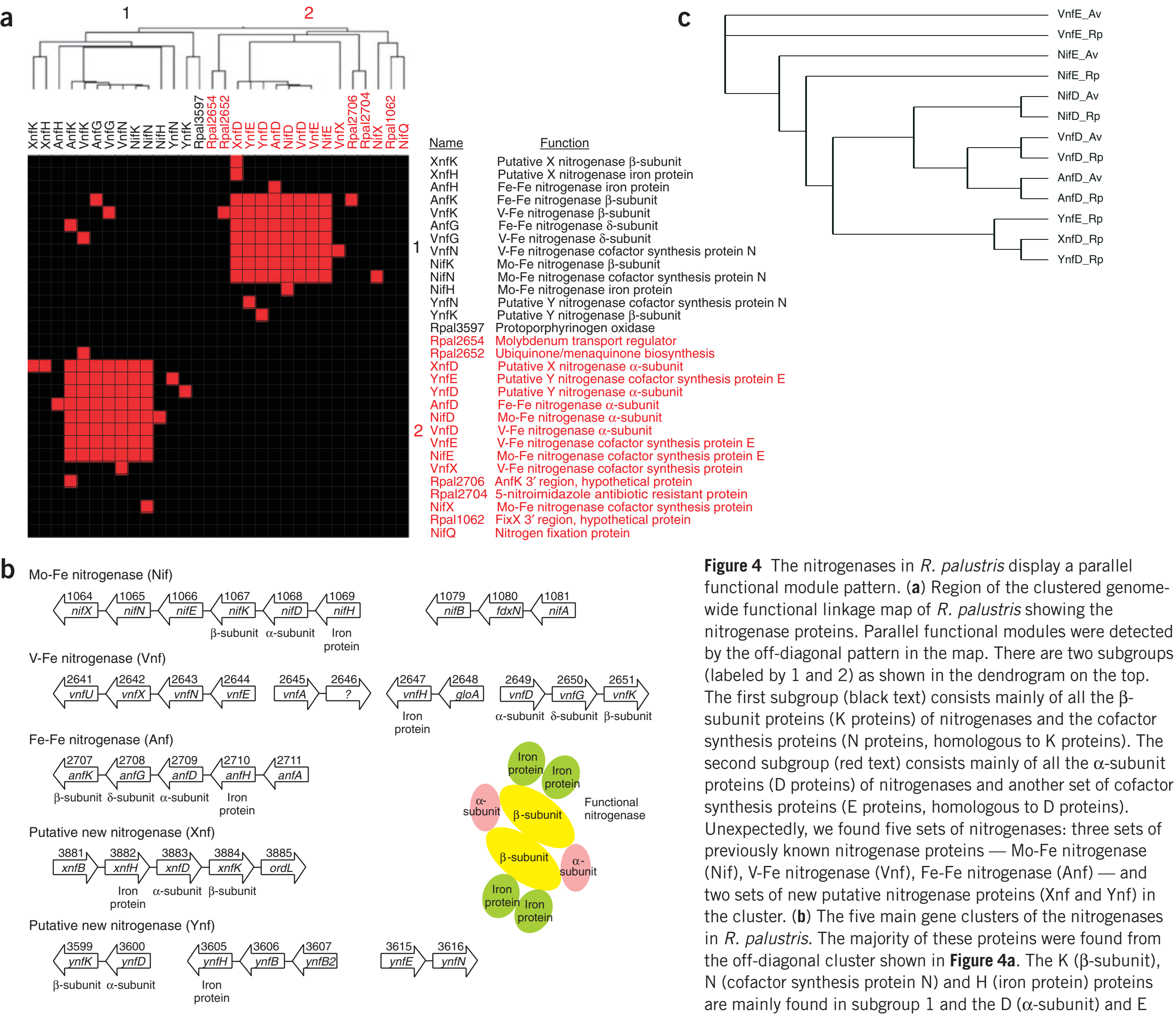
Discovery of parallel functional modules

Our analysis of ten genomes suggests that a given set of parallel functional modules may be specific to a given organism, reflecting the life style of the organism. For example, *R. palustris* has a metabolically versatile life style[36]. It survives in diverse environments; it fixes atmospheric nitrogen into ammonia; it uses carbon dioxide gas and various aromatic compounds as its carbon sources. Our parallel functional module analysis shows that these aspects of the organism are reflected in its genome as multiple parallel pathways or complexes of transporters, multidrug efflux pumps, nitrogenases, nitrogen utilization enzymes and aromatic compound degradation enzymes (**Table 1**). Another example comes from a



	Name	Function	Cellular localization
	DppB	Dipeptide transport system permease protein 1	Inner membrane
	NikB	Transport of nickel, membrane protein	Inner membrane
	OppB	Oligopeptide transport permease protein	Inner membrane
1	B0831	Putative transport system permease protein	Inner membrane
	B1486	Putative transport system permease protein	Inner membrane
	SapB	Peptide transport permease protein	Inner membrane
	YejB	Putative transport system permease protein	Inner membrane
	NikE	ATP-binding protein of nickel transport system	Cytoplasm
	B1483	Putative ATP-binding component of a transport system	Cytoplasm
2	OppF	ATP-binding protein of oligo peptide ABC transport system	Cytoplasm
	DppF	Putative ATP-binding component of dipeptide transport system	Cytoplasm
	SapF	Putative ATP-binding protein of peptide transport system	Cytoplasm
	YejF	Putative ATP-binding component of a transport system	Inner membrane
	NikD	ATP-binding protein of nickel transport system	Inner membrane
3	B1484	Putative ATP-binding component of a transport system	Cytoplasm
	OppD	ATP-binding protein of oligo peptide ABC transport system	Inner membrane
	DppD	Putative ATP-binding component of dipeptide transport system	Inner membrane
	SapD	Putative ATP-binding protein of peptide transport system	Cytoplasm

Figure 3 The peptide transporters in *E. coli* K12 display a parallel complex pattern. (**a**) The clustered genome-wide functional linkage map of *E. coli* K12 showing peptide transporters. Three known peptide transporters in the *E. coli* genome (Opp, Dpp and Sap), the nickel transporter (Nik), and three sets of uncharacterized proteins (B1483-B1487, B0829-B0832, YejABEF) were clustered together. Three subgroups corresponding to the three components of the transporters are shown here as indicated by numbers 1–3 and by the dendrogram on the top. The first subgroup (black text) is composed of mostly permease proteins of the seven transporters. The second (red text) and the third (cyan text) subgroups are composed of the ATP-binding proteins of the transporters. On the right, the cellular location of each protein predicted by PSORT[33] is shown. Most of the proteins in the same subgroups are predicted to have the same cellular locations. (**b**) The functional linkage networks of the peptide transporters in *E. coli* derived by the four-step approach. Five (Opp, Dpp, Sap, Nik and B1483-B1487) of the seven parallel complexes shown here have the same network structures. We infer that the previously uncharacterized proteins (B1483-B1487, B0829-B0832, YejABEF) in the cluster function as peptide transporters, too. Notice that the upper left corner of the functional linkage map contains some red squares, which are not consistent with our ideal schematic off-diagonal cluster pattern. These show that the proteins DppB, NikB, OppB and B0831 are functionally linked, although they are in the same subgroup. This effect arises from the coevolution linkages among the four proteins. These extra linkages are removed in step 4.



Name	Function
XnfK	Putative X nitrogenase β-subunit
XnfH	Putative X nitrogenase iron protein
AnfH	Fe-Fe nitrogenase iron protein
AnfK	Fe-Fe nitrogenase β-subunit
VnfK	V-Fe nitrogenase β-subunit
AnfG	Fe-Fe nitrogenase δ-subunit
VnfG	V-Fe nitrogenase δ-subunit
VnfN	V-Fe nitrogenase cofactor synthesis protein N
NifK	Mo-Fe nitrogenase β-subunit
NifN	Mo-Fe nitrogenase cofactor synthesis protein N
NifH	Mo-Fe nitrogenase iron protein
YnfN	Putative Y nitrogenase cofactor synthesis protein N
YnfK	Putative Y nitrogenase β-subunit
Rpal3597	Protoporphyrinogen oxidase
Rpal2654	Molybdenum transport regulator
Rpal2652	Ubiquinone/menaquinone biosynthesis
XnfD	Putative X nitrogenase α-subunit
YnfE	Putative Y nitrogenase cofactor synthesis protein E
YnfD	Putative Y nitrogenase α-subunit
AnfD	Fe-Fe nitrogenase α-subunit
NifD	Mo-Fe nitrogenase α-subunit
VnfD	V-Fe nitrogenase α-subunit
VnfE	V-Fe nitrogenase cofactor synthesis protein E
NifE	Mo-Fe nitrogenase cofactor synthesis protein E
VnfX	V-Fe nitrogenase cofactor synthesis protein
Rpal2706	AnfK 3′ region, hypothetical protein
Rpal2704	5-nitroimidazole antibiotic resistant protein
NifX	Mo-Fe nitrogenase cofactor synthesis protein
Rpal1062	FixX 3′ region, hypothetical protein
NifQ	Nitrogen fixation protein

Figure 4 The nitrogenases in *R. palustris* display a parallel functional module pattern. (**a**) Region of the clustered genome-wide functional linkage map of *R. palustris* showing the nitrogenase proteins. Parallel functional modules were detected by the off-diagonal pattern in the map. There are two subgroups (labeled by 1 and 2) as shown in the dendrogram on the top. The first subgroup (black text) consists mainly of all the β-subunit proteins (K proteins) of nitrogenases and the cofactor synthesis proteins (N proteins, homologous to K proteins). The second subgroup (red text) consists mainly of all the α-subunit proteins (D proteins) of nitrogenases and another set of cofactor synthesis proteins (E proteins, homologous to D proteins). Unexpectedly, we found five sets of nitrogenases: three sets of previously known nitrogenase proteins — Mo-Fe nitrogenase (Nif), V-Fe nitrogenase (Vnf), Fe-Fe nitrogenase (Anf) — and two sets of new putative nitrogenase proteins (Xnf and Ynf) in the cluster. (**b**) The five main gene clusters of the nitrogenases in *R. palustris*. The majority of these proteins were found from the off-diagonal cluster shown in **Figure 4a**. The K (β-subunit), N (cofactor synthesis protein N) and H (iron protein) proteins are mainly found in subgroup 1 and the D (α-subunit) and E (cofactor synthesis protein E) proteins are mainly found in subgroup 2. Right arrows indicate the genes encoded on the coding strand of the DNA. Left arrows indicate the genes encoded on the complementary strand of the DNA. The numbers above the arrows indicate the positions of the genes on the chromosome based on the genome sequence released on the NCBI website, version NZ_AAAF01000001.1 (9/24/2003). (**c**) The protein tree of the α-subunits and the cofactor synthesis proteins E of the nitrogenases from *R. palustris* and *A. vinelandii* constructed from sequence alignment. As expected, the NifD proteins from *R. palustris* (NifD_Rp) and *A. vinelandii* (NifD_Av) are clustered most closely with each other. The same is true for the VnfD proteins and AnfD proteins. The α-subunits of the two new putative nitrogenases from *R. palustris* (XnfD_Rp and YnfD_Rp) are clustered together and are distantly related to the α-subunits of the three known nitrogenases. Nif, Mo-Fe nitrogenase; Vnf, V-Fe nitrogenase; Anf, Fe-Fe nitrogenase; Xnf, putative new nitrogenase; Ynf, putative new nitrogenase; D, α-subunit of nitrogenase; E, cofactor synthesis protein E; Rp, *R. palustris*; Av, *A. vinelandii*.

thermophile *P. aerophilum*, which lives in high-temperature and oxidative environments. Its genome encodes two parallel ferredoxin oxidoreductase complexes that were revealed in our analysis (**Table 1**). On the other hand, some parallel functional modules are common to many organisms, possibly because the functions of those modules are essential for survival. For example, heat shock proteins have multiple parallel complexes in all the eukaryotic genomes that we examined.

Identifying parallel functional modules can help to interpret the physiology of an organism from its genome sequences. The existence of parallel functional modules for a given function may suggest that the organism is versatile in performing the function and could better survive in diverse environments where the function is needed. For example, we found six peptide transporters in *E. coli* K12. Bacteria use peptides as carbon and nitrogen sources[28]. Peptide transporters have distinct substrate specificities and spectra[39,40]. Different peptide transporters may be needed for *E. coli* to survive in various environments, where available nutrients can be different. In the case of *R. palustris*, the organism has four parallel oxygenase-dependent ring cleavage pathways with different substrate specificities[36]. Thus *R. palustris* uses various aromatic compounds as carbon sources and survives in different environments.

Other approaches to discover parallel functional modules

Another approach to discover parallel functional modules is that of Kelley *et al*[5]., described in the introduction, which uses large-scale experimental protein interaction data. Another possible method, which uses genome sequences, as our method does, is to combine homology searches with gene location relationships. If the components of a functional module are known a priori one may identify its parallel modules using traditional sequence alignment. One can search for paralogs of the target proteins in the genome and then extend the parallel functional modules by including the operon partners of the paralogs. This rudimentary approach has limitations. First, it needs to have predetermined target functional modules before searching for parallel modules. Second, ancient gene duplication events make it difficult to reliably detect homology among genes. Third, this homology-based approach does not provide much information on the network structure of the functional modules. Fourth, it is difficult to apply this approach to eukaryotic genomes without a preexisting list of all the members in a functional module, because operon structure cannot be used for finding other members in the parallel modules.

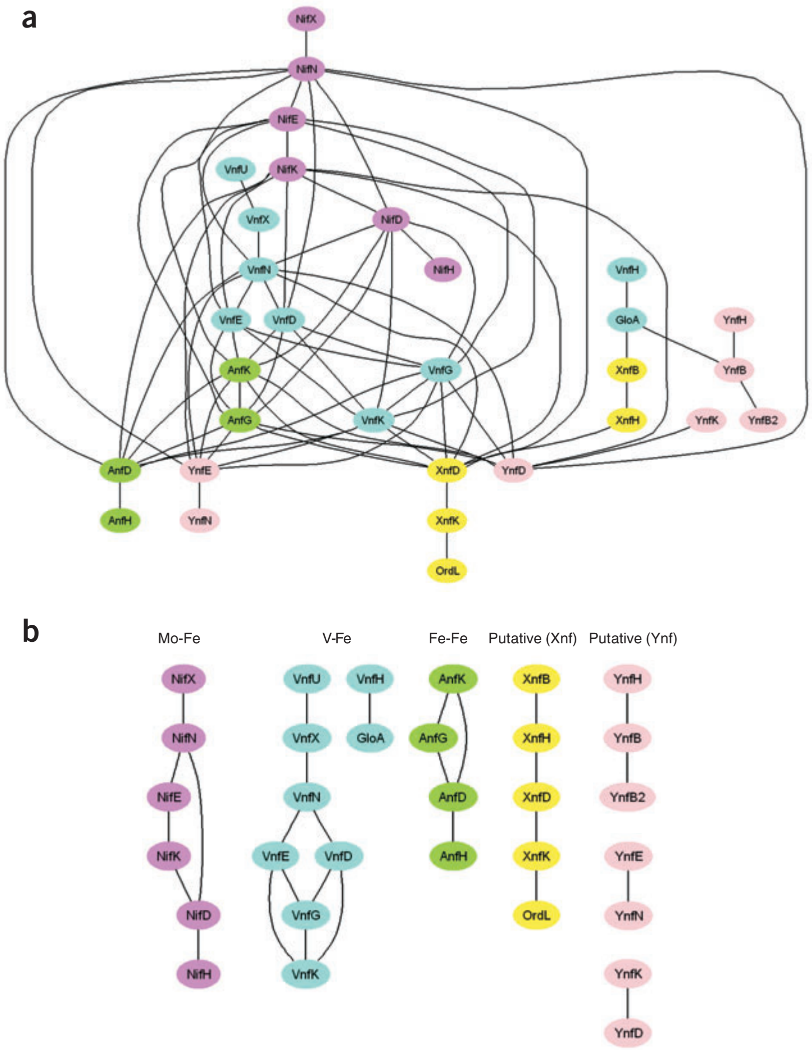


Figure 5 The functional linkage networks of nitrogenase proteins in *R. palustris* before and after untangling the parallel functional modules. (**a**) Before untangling the parallel functional modules, the relationships among the nitrogenase proteins are depicted as a massive network with many spurious linkages among the proteins in different functional modules. (**b**) After untangling the parallel functional modules and removing the spurious linkages, the relationships among the nitrogenase modules are depicted clearly, including the protein components in each module and their paralogs in other modules.

Features of the four-step approach

The four-step approach offers genome-wide discovery of parallel functional modules. It is a discovery-driven approach unrestrained by the need to focus on a predetermined target. The approach uses genome sequences. Thus, it can be applied to all fully sequenced organisms and is not limited by the availability of experimental interaction data. The approach is also able to identify the parallel functional modules that are encoded in the genomes but may not be expressed under the experimental conditions. Parallel functional modules usually have partially redundant functions. The redundant genes may be expressed only under specific conditions and hence their expression may not be represented in most experimental data.

The four-step approach not only discovers parallel functional modules, but does so in the context of inferred protein networks, simultaneously revealing the functional relationships among the proteins within modules. The functions of proteins depend on their biological context—their relations to other proteins and the set of interactions that they form[15]. The inference methods functionally link proteins that in general do not have sequence similarity. The context and connectivity of the interactions inferred from the approach add information about the functions of the proteins.

The four-step approach can provide higher-resolution inference of protein functions, based on the membership of the proteins in subgroups, than can homology-based methods. For example, two cofactor synthesis proteins E and N from the same nitrogenase share marginal sequence similarity. Their paralogs from different nitrogenases are not distinguishable because of low sequence similarity to their ancestor proteins. Thus it is difficult to assign the function of cofactor synthesis proteins E or N. The four-step approach sorted two cofactor synthesis proteins E and N of Mo-Fe, V-Fe and Ynf nitrogenases into two separate subgroups (**Fig. 4a**). Therefore, it distinguished the E from the N proteins of the new putative nitrogenase Ynf based on the membership of the known nitrogenase proteins. In contrast, the traditional homology-based method could not separate the two cofactor synthesis proteins for the new putative nitrogenase.

At present, protein functional linkages in eukaryotic genomes are mainly inferred based on the protein homologs in bacterial genomes. The number of linkages is limited by the available homologs in prokaryotes. Thus, eukaryote-specific functional modules were not revealed in this study. As more eukaryotic genome sequences become available, we expect that more functional linkages will be inferred from comparison of eukaryotic genomes. It is more difficult to pair the functional module partners from the subgroups in eukaryotic genomes than in prokaryotic genomes owing to the lack of conservation in gene order. But in step 4, which untangles the functional linkages among parallel functional modules, not only can the phylogenetic distance matrices method be used, but also the interacting protein pairs deduced from large-scale experimental data, which are more readily available for eukaryotic genomes. Such data includes cellular colocalization, common transcription regulators and *cis*-elements of genes, gene coexpression and synthetic lethal analysis.

In summary, our comparative analysis of genome sequences using the four-step approach uncovers parallel functional modules on a genomic scale. The approach reveals the functional relationships among the proteins within the modules and provides higher-resolution inference of protein functions and interactions.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS
We thank Thomas G. Graeber and Michael J. Thompson for helpful discussions and HHMI and Department of Energy for support.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/

1. Henikoff, S. *et al.* Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**, 609–614 (1997).
2. Teichmann, S.A., Park, J. & Chothia, C. Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proc. Natl. Acad. Sci. USA* **95**, 14658–14663 (1998).
3. Brenner, S.E., Hubbard, T., Murzin, A. & Chothia, C. Gene duplications in H. influenzae. *Nature* **378**, 140 (1995).
4. Gough, J., Karplus, K., Hughey, R. & Chothia, C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919 (2001).
5. Kelley, B.P. *et al.* Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci. USA* **100**, 11394–11399 (2003).
6. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
7. Huynen, M.A., Snel, B., von Mering, C. & Bork, P. Function prediction and protein networks. *Curr. Opin. Cell Biol.* **15**, 191–198 (2003).
8. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288 (1999).
9. Marcotte, E.M. *et al.* Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
10. Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
11. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328 (1998).
12. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **96**, 2896–2901 (1999).
13. Pellegrini, M., Thompson, M., Fierro, J. & Bowers, P. Computational method to assign microbial genes to pathways. *J. Cell. Biochem. Suppl.* **37**, 106–109 (2001).
14. Bowers, P.M. *et al.* Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* **5**, R35 (2004).
15. Eisenberg, D., Marcotte, E.M., Xenarios, I. & Yeates, T.O. Protein function in the post-genomic era. *Nature* **405**, 823–826 (2000).
16. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
17. Strong, M. *et al.* Visualization and interpretation of protein networks in Mycobacterium tuberculosis based on hierarchical clustering of genome-wide functional linkage maps. *Nucleic Acids Res.* **31**, 7099–7109 (2003).
18. Ramani, A.K. & Marcotte, E.M. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.* **327**, 273–284 (2003).
19. Gertz, J. *et al.* Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics* **19**, 2039–2045 (2003).
20. Baumann, P., Qureshi, S.A. & Jackson, S.P. Transcription: new insights from studies on Archaea. *Trends Genet.* **11**, 279–283 (1995).
21. Archambault, J. & Friesen, J.D. Genetics of eukaryotic RNA polymerases I, II, and III. *Microbiol. Rev.* **57**, 703–724 (1993).
22. Falkenburg, D., Dworniczak, B., Faust, D.M. & Bautz, E.K. RNA polymerase II of *Drosophila*. Relation of its 140,000 Mr subunit to the beta subunit of *Escherichia coli* RNA polymerase. *J. Mol. Biol.* **195**, 929–937 (1987).
23. Seifarth, W. *et al.* Identification of the genes coding for the second-largest subunits of RNA polymerases I and III of *Drosophila melanogaster. Mol. Gen. Genet.* **228**, 424–432 (1991).
24. Knackmuss, S., Bautz, E.F. & Petersen, G. Identification of the gene coding for the largest subunit of RNA polymerase I (A) of *Drosophila melanogaster. Mol. Gen. Genet.* **253**, 529–534 (1997).
25. Jokerst, R.S., Weeks, J.R., Zehring, W.A. & Greenleaf, A.L. Analysis of the gene encoding the largest subunit of RNA polymerase II in *Drosophila. Mol. Gen. Genet.* **215**, 266–275 (1989).
26. The FlyBase database of the. *Drosophila* genome projects and community literature. *Nucleic Acids Res.* **31**, 172–175 (2003).
27. Linton, K.J. & Higgins, C.F. The *Escherichia coli* ATP-binding cassette (ABC) proteins. *Mol. Microbiol.* **28**, 5–13 (1998).
28. Hogarth, B.G. & Higgins, C.F. Genetic organization of the oligopeptide permease (opp) locus of *Salmonella typhimurium* and *Escherichia coli. J. Bacteriol.* **153**, 1548–1551 (1983).
29. Smith, M.W., Tyreman, D.R., Payne, G.M., Marshall, N.J. & Payne, J.W. Substrate specificity of the periplasmic dipeptide-binding protein from *Escherichia coli*: experimental basis for the design of peptide prodrugs. *Microbiology* **145**, 2891–2901 (1999).
30. Parra-Lopez, C., Baer, M.T. & Groisman, E.A. Molecular genetic analysis of a locus required for resistance to antimicrobial peptides in *Salmonella typhimurium. EMBO J.* **12**, 4053–4062 (1993).
31. Chen, H.Y., Weng, S.F. & Lin, J.W. Identification and analysis of the *sap* genes from Vibrio fischeri belonging to the ATP-binding cassette gene family required for peptide transport and resistance to antimicrobial peptides. *Biochem. Biophys. Res. Commun.* **269**, 743–748 (2000).
32. Navarro, C., Wu, L.F. & Mandrand-Berthelot, M.A. The nik operon of *Escherichia coli* encodes a periplasmic binding-protein-dependent transport system for nickel. *Mol. Microbiol.* **9**, 1181–1191 (1993).
33. Nakai, K. & Horton, P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* **24**, 34–36 (1999).
34. Eady, R.R. Structure-function relationships of alternative nitrogenases. *Chem. Rev.* **96**, 3013–3030 (1996).
35. Howard, J.B. & Rees, D.C. Structural basis of biological nitrogen fixation. *Chem. Rev.* **96**, 2965–2982 (1996).
36. Larimer, F.W. *et al.* Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris. Nat. Biotechnol.* **22**, 55–61 (2004).
37. Brigle, K.E., Weiss, M.C., Newton, W.E. & Dean, D.R. Products of the iron-molybdenum cofactor-specific biosynthetic genes, nifE and nifN, are structurally homologous to the products of the nitrogenase molybdenum-iron protein genes, nifD and nifK. *J. Bacteriol.* **169**, 1547–1553 (1987).
38. Wolfinger, E.D. & Bishop, P.E. Nucleotide sequence and mutational analysis of the vnfENX region of *Azotobacter vinelandii. J. Bacteriol.* **173**, 7565–7572 (1991).
39. Hagting, A., Kunji, E.R., Leenhouts, K.J., Poolman, B. & Konings, W.N. The di- and tripeptide transport protein of *Lactococcus lactis*. A new type of bacterial peptide transporter. *J. Biol. Chem.* **269**, 11391–11399 (1994).
40. Higgins, C.F. & Gibson, M.M. Peptide transport in bacteria. *Methods Enzymol.* **125**, 365–377 (1986).
41. Gur, E. *et al.* The *Escherichia coli* DjlA and CbpA proteins can substitute for DnaJ in DnaK-mediated protein disaggregation. *J. Bacteriol.* **186**, 7236–7242 (2004).
42. Kaneko, T. *et al.* Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. *DNA Res.* **9**, 189–197 (2002).
43. Gelfand, M.S. & Koonin, E.V. Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Res.* **25**, 2430–2439 (1997).
44. Morano, K.A., Liu, P.C. & Thiele, D.J. Protein chaperones and the heat shock response in *Saccharomyces cerevisiae. Curr. Opin. Microbiol.* **1**, 197–203 (1998).
45. Piekny, A.J., Johnson, J.L., Cham, G.D. & Mains, P.E. The *Caenorhabditis elegans* nonmuscle myosin genes nmy-1 and nmy-2 function as redundant components of the let-502/Rho-binding kinase and mel-11/myosin phosphatase pathway during embryonic morphogenesis. *Development* **130**, 5695–5704 (2003).
46. Price, N.T., Jackson, V.N. & Halestrap, A.P. Cloning and sequencing of four new mammalian monocarboxylate transporter (MCT) homologues confirms the existence of a transporter family with an ancient past. *Biochem. J.* **329**, 321–328 (1998).
47. Heschl, M.F. & Baillie, D.L. The HSP70 multigene family of *Caenorhabditis elegans. Comp. Biochem. Physiol. B* **96**, 633–637 (1990).
48. Bettencourt, B.R. & Feder, M.E. Rapid concerted evolution via gene conversion at the *Drosophila* hsp70 genes. *J. Mol. Evol.* **54**, 569–586 (2002).
49. Bettencourt, B.R. & Feder, M.E. Hsp70 duplication in the *Drosophila melanogaster* species group: how and when did two become five? *Mol. Biol. Evol.* **18**, 1272–1282 (2001).
50. Borner, G.H., Sherrier, D.J., Stevens, T.J., Arkin, I.T. & Dupree, P. Prediction of glycosylphosphatidylinositol-anchored proteins in *Arabidopsis*. A genomic analysis. *Plant Physiol.* **129**, 486–499 (2002).
51. Lin, B.L. *et al.* Genomic analysis of the Hsp70 superfamily in *Arabidopsis thaliana. Cell Stress Chaperones* **6**, 201–208 (2001).