# Computational genetics: finding protein function by nonhomology methods

Edward M Marcotte

During the past year, computational methods have been developed that use the rapidly accumulating genomic data to discover protein function. The methods rely on properties shared by functionally related proteins other than sequence or structural similarity. Instead, these 'nonhomology' methods analyze patterns such as domain fusion, conserved gene position and gene co-inheritance and coexpression to identify protein–protein relationships. The methods can identify functions for proteins that are without characterized homologs and have been applied to genome-wide predictions of protein function.

**Addresses**
Molecular Biology Institute, UCLA-DOE Laboratory of Structural Biology and Molecular Medicine, University of California Los Angeles, PO Box 951570, Los Angeles, CA 90095-1570, USA and Protein Pathways Inc., 1145 Gayley Avenue, Ste 304, Los Angeles, CA 90024, USA; e-mail: marcotte@mbi.ucla.edu

**Current Opinion in Structural Biology** 2000, **10**:359–365

0959-440X/00/$ – see front matter
© 2000 Elsevier Science Ltd. All rights reserved.

**Abbreviations**
**COGS** clusters of orthologs
**EST** expressed sequence tag

## Introduction

Biologists are in a delightful quandary. Thousands of potential genes are being discovered in the various genome sequencing projects, including those encoding many new families of proteins. Often, these proteins are evolutionarily conserved, but are of unknown function. This poses a fundamental problem to biologists: how can we discover the functions of these thousands of unknown proteins quickly and efficiently? Even more ambitious than knowing their specific biochemical functions, can we discover their broader functions — the cellular context, such as pathways and complexes, in which they operate?

As difficult as this goal is, significant progress has been made in the past year both experimentally, by conducting genome-wide experiments measuring, for example, mRNA expression [1] or biochemical activity [2•], and computationally, by developing new analyses that work on fundamentally different principles from homology- or structure-based methods.

This *in silico* progress stemmed from the realization that genomes contain considerable information about the functions of and relationships between genes and proteins. This functional information is encoded in forms such as patterns of gene fusion, conservation of gene position, patterns of gene co-inheritance and other sorts of evolutionary information. Such patterns are revealed by comparisons of multiple genomes, making these analyses only recently tractable. Also, additional data, such as gene coexpression measurements, provide analogous information within single organisms.

The power of these new methods is that they produce networks of functionally related proteins, even when the proteins have never been characterized. Protein function is defined by these methods in terms of context, that is, which cellular pathways or complexes the protein participates in, rather than by suggesting a specific biochemical activity. However, in cases in which some of the proteins have a known function, their function can be extended to the most intimately linked uncharacterized proteins. Thus, the methods can be used both to find functional relationships and to assign general protein function.

This results in an approach to finding protein function that is strikingly different from directly comparing amino acid sequences, although sequence comparisons are the basic tool used in many of the methods. The functional information discovered also differs from what might be learned either from direct sequence comparisons or from structural analyses, giving three relatively independent and complementary routes to protein function, as shown in Figure 1. This review will discuss the main ideas behind nonhomology methods, the newest route to protein function.

## Evolution (some homology required)

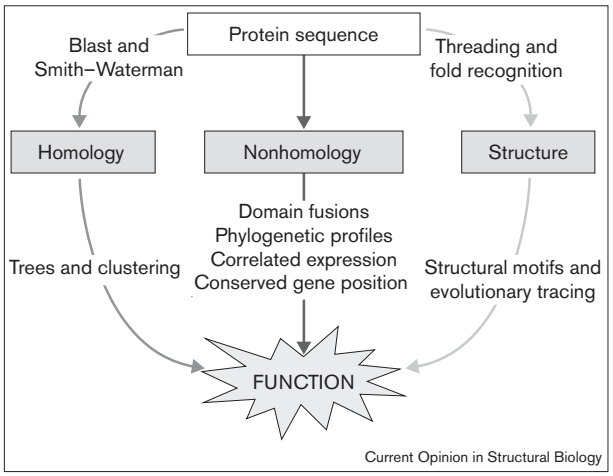
Several nonhomology methods take advantage of genetic variations among organisms to find protein function. The *domain fusion method* [3•] finds functionally related proteins by analyzing patterns of domain fusion. As illustrated in Figure 2, proteins found separately in one organism can often be found fused into a single polypeptide chain in another organism. That the separated proteins have a functional relationship can be inferred from knowledge of the fused protein, named the Rosetta stone protein for its ability to reveal the relationship among its component parts.

In many cases, the proteins linked by such a domain fusion event may even physically interact, especially in the case of protein pairs that have been filtered for false-positive-producing 'promiscuous domains' [3•] and in cases of high-scoring sequence matches to the Rosetta stone protein [4•]. An example of this is the two *Escherichia coli* gyrase subunits GyrA and GyrB, which are found as fused homologs in yeast topoisomerase II [3•]. Such relationships are also common among separated eukaryotic proteins found fused in a prokaryotic Rosetta stone protein (EM Marcotte, unpublished data).

In its simplest form, this analysis can be implemented [3•] by searching a large sequence database for homologs

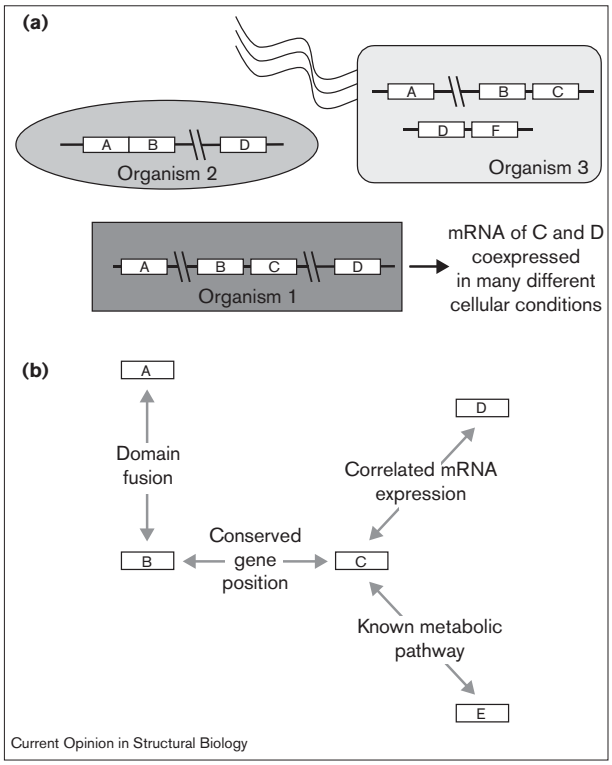
DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

**Figure 1**




One can take several computational routes to discovering the function of a protein. On the left-hand route, the protein sequence is compared directly with other protein sequences [44,45]. Characterized sequence homologs or phylogenetic analyses (as in [17–19]) may suggest functional information. On the right-hand route, the protein sequence may be tested for compatibility with known three-dimensional protein structures [46]. Knowledge of the structure may then suggest functional information (e.g. as in [47,48]). Along the middle route are nonhomology methods. Sequence and structural homology reveal proteins of identical or equivalent function, whereas nonhomology methods identify interacting proteins, proteins with related functions or proteins operating in the same cellular context. Nonhomology methods return a network of relationships among proteins functionally linked to the query protein and function is both defined and inferred by this network of related proteins.

**Figure 2**




An example of deriving protein–protein relationships by nonhomology methods. Genes (labeled white boxes) are shown on the chromosomes (thick horizontal lines) of three different organisms. **(a)** It can be inferred that the proteins encoded by genes **A–D** are functionally related through patterns such as the conserved gene positions of **B** and **C** in organisms 1 and 3, the fusion of **A** and **B** into **A–B** in organism 2 and the coexpression of the mRNAs of **C** and **D** in organism 1. These results can be represented as a network of functional relationships, as shown in **(b)**. If, for example, the function of **B** was unknown, it might be inferred from the functions of proteins **A** and **C**. The computational linkages may be supplemented by any experimentally observed interactions or known protein–protein relationships, such as those described in [35,38–40].

of a query protein **A**. Hits in this search will include direct homologs of the query protein (**A′**) and potential Rosetta stone fusion proteins (**A–B**). Each hit is then used as a query to search the genome of **A** and functionally related **B** proteins will be found in this second search. Along with **B** proteins, hits in this search will include **A**, homologs of **A** (**A′**) and very distant homologs of **A** (**A″**) [5], but the **B** proteins can be identified by their lack of homology to **A** or by their homology to different regions of **A–B** than those homologous to **A**. Such an analysis recently proved useful in identifying a functional relationship between CHORD-containing proteins and Sgt1, proteins important for plant disease signaling and nematode development [6].

In a related fashion, two proteins can be inferred to be functionally related if their genes are repeatedly found as neighbors on the chromosomes of different organisms [7•,8,9•], as shown in Figure 2. This *conservation of relative gene position* presumably derives from the organization of prokaryotic genes into operons in which each protein encoded by the operon performs a closely related task, such as the proteins of the lactose system [10] or proteins involved in iron uptake [11]. To find operons directly would require the identification of promoters and regulatory elements; however, for operons containing evolutionarily conserved genes, large portions of the operons can often be reconstructed automatically simply by identifying pairs of conserved gene neighbors [9•].

As with domain fusion analysis, this approach can often identify interacting proteins [7•]. How well the approach will extend to eukaryotic genes remains to be seen, as eukaryotes generally lack operons. Examples of functionally related eukaryotic gene neighbors do exist, however, such as in the TCL1 locus [12] or the cadherin proteins [13], so the technique may be useful. The quality of the functional relationships identified by this method is exceptional, but the coverage is unfortunately low because of the dual requirement of identifying orthologs in another genome and then finding those orthologs that are adjacent on the chromosome.

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

A third nonhomology method works on the premise that proteins that operate together in the cell are often inherited in a correlated fashion [14•]. That this is a reasonable assumption follows from the fact that proteins rarely work alone and many pathways or complexes are crippled by the loss of individual components. Thus, any organism that requires the complex or pathway carries the genes for most or all of its components; any organism lacking the complex or pathway often lacks all of the component genes. The co-inherited proteins can be identified in an automated fashion by comparing their *phylogenetic profiles*, strings that encode the presence or absence of sequence homologs in known genomes, as shown for a few proteins in Figure 3.

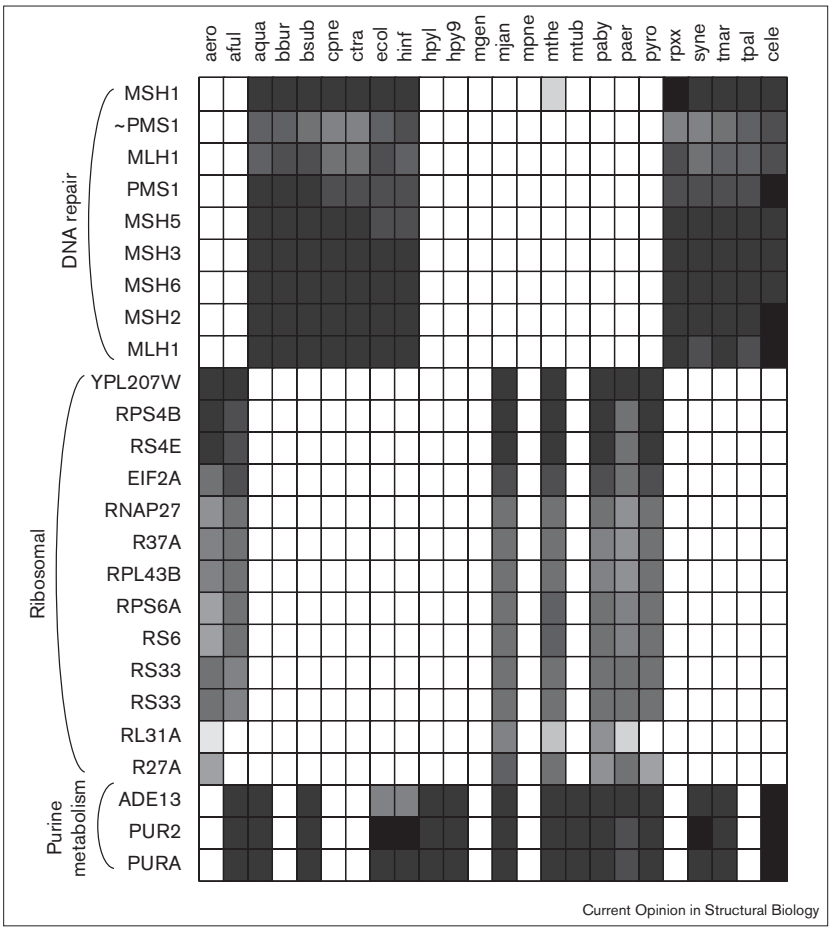
Each phylogenetic profile is analogous to an abstract representation of an evolutionary tree; matching phylogenetic profiles therefore identifies proteins with similar patterns of inheritance. Note that no homology is required among the proteins with similar phylogenetic profiles; the proteins are co-inherited and, when many genomes ($n > 10$) are analyzed, usually functionally related, as in the examples shown in Figure 3. The involvement of the uncharacterized SmpB protein family in protein synthesis, predicted by phylogenetic profiles [14•], was recently confirmed by Karzai *et al.* [15].

The *differential genome analysis* method of Huynen *et al.* [16] also takes advantage of gene presence and absence to associate phenotypes with genes: a list is prepared of genes shared among organisms that also share a given phenotype. This list of genes is filtered by removing the genes that occur in organisms lacking the phenotype. The remaining genes are correspondingly enriched for those that confer the phenotype.

## Homology (and evolution)

The distinction between homology and nonhomology methods can be blurred, as even direct sequence comparisons are enhanced by taking advantage of evolutionary variations. For example, Lichtarg *et al.* [17] showed that functional sites on proteins could be identified by analyzing amino acids conserved at different branching depths in phylogenetic trees of protein homologs. Likewise, variations among protein homologs found by clustering the proteins in phylogenetic trees often reveal subtle specialization in protein function. Recent examples of this

**Figure 3**




Phylogenetic profiles [14•] for three groups of yeast proteins (ribosomal proteins and proteins involved in DNA repair and purine metabolism) sharing similar co-inheritance patterns. Each row is a graphical representation of a protein phylogenetic profile, with elements colored according to whether a homolog is absent (white box) or present (colored box) in each of 24 genomes (columns). When homology is present, the elements are shaded on a gradient from light gray (low homology) to black (strong homology). In this case, homologs are considered absent when no BLAST hits [44] are found with expectation ($E$) values $< 1 \times 10^{-5}$. When homologs are present, the profile receives a score ($-1/\log E$) that describes the degree of sequence similarity with the best match in that genome. Note that an uncharacterized protein (YPL207W) clusters with the ribosomal proteins and can now be assigned a function in protein synthesis.

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

analysis include the MutS protein family [18] and proteins conserved between worm and yeast [19].

A different aspect of evolutionary information is used in the calculation of COGS (clusters of orthologs) [20], in which proteins from different organisms are grouped together in such a way as to maximize their functional equivalence. COGS are generated by identifying orthologs or equivalent proteins among different organisms. Orthologs can be defined operationally as the symmetric top-scoring protein sequences in a sequence homology search. That is, a query sequence from genome 1 has an ortholog in genome 2 if searching the query versus genome 2 turns up the ortholog as the best match and searching the ortholog versus genome 1 turns up the query protein as the best match.

COGS effectively cluster functionally equivalent proteins because of the power of orthology, which dictates that not only are sequences homologous, but also that they are the best homologs regardless of search direction. This symmetric homology detection relies on the absence of better homologs from each genome and, therefore, incorporates both evolutionary information and sequence matching. Phylogenetic profiles can be constructed from orthologs, rather than from best homologs, and searched for exact matches at the COGS web site (http://www.ncbi.nlm.nih.gov/COG/) [20].

## No homology required (made up for with extra data)

Each of the methods discussed above requires that a query protein have some sequence homologs in the database, even though direct sequence homology with these proteins may not be the basis for the analysis. This requirement is lifted for analyses of other genomic data, however, such as analysis of *correlated mRNA expression* levels, reviewed in [21,22]. Therefore, these techniques can find relationships among proteins that are absolutely unique. The premise of all expression clustering methods is, as in phylogenetic profiles, that proteins rarely work alone, but are often expressed at the same time or place as functionally related proteins. By varying the conditions that cells are grown in or by choosing different cell types or cells from different tissues, enough variation in gene expression can be observed to identify coexpressing genes.

Such clustering requires additional data beyond genome sequences, to date relying on measurements of cellular mRNA levels by DNA microarrays, as in [23], serial analysis of gene expression (SAGE) libraries [24] or expressed sequence tag (EST) libraries [25]. Underlying the clustering of genes by their mRNA coexpression levels is the assumption that coexpressed genes will generally be functionally related if enough different conditions have been tested. Such clustering performs well for strongly coexpressed genes, such as ribosomal subunits, and poorly for other gene groups. It requires fairly large sets of data, such as more than 70 DNA chip measurements of yeast mRNA levels [26••,27••] or hundreds of human EST libraries from different tissues and cells [28•].

In a manner analogous to analyzing gene co-inheritance or mRNA expression patterns, an organism's proteins can probably be clustered effectively by their own protein coexpression patterns under varying growth conditions. For *protein coexpression analysis*, one directly measures the functional species (proteins) and it is likely that the clusters calculated on this basis will be more powerful for protein function assignment than mRNA expression clustering, especially given that protein and mRNA levels are often surprisingly uncorrelated [29•]. Protein expression patterns have been measured directly by mass spectrometry of protein mixtures [30] and by various two-dimensional gel electrophoresis techniques, with the proteins on the gel identified by amino acid content [31] or mass spectrometry [32•,33•]. The techniques are labor intensive and have not, just yet, produced a sufficient dataset for coexpression analyses. This is likely to change in the very near future, given the current emphasis on genome-wide and proteome-wide analyses. Protein expression patterns have also been measured, not as a function of growth conditions, but spatially, as β-galactosidase fusions in *Xenopus* embryos [34•], allowing functionally related proteins to be grouped by their spatial coexpression patterns.

## Building a genome-wide network of proteins

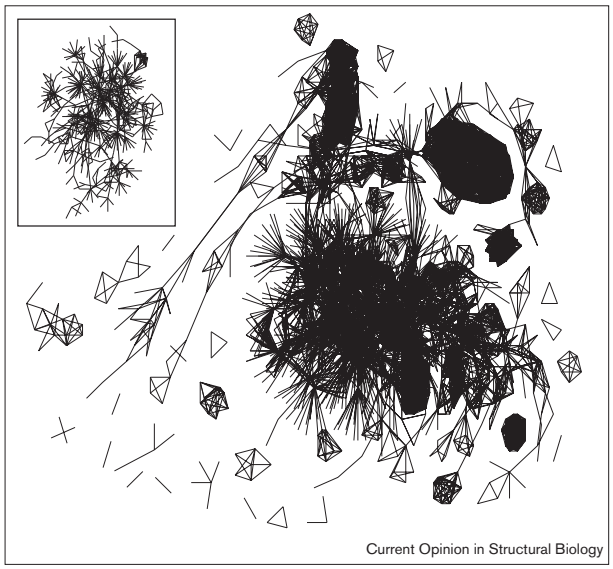
The methods described above are easily applied on a genome-wide scale, combining results from each method to build a network of the functional relationships among an organism's proteins. Such a network was calculated recently for yeast proteins [27••], identifying 93,750 functional links among 4701 of the 6217 proteins in yeast. A subset of this network is drawn in Figure 4, showing the amazing complexity of the connections generated by these methods. Perhaps even more surprising is the high degree of connectivity among the proteins, attributable in part to homology and false-positive predictions, but still observable even in entirely experimentally derived networks, such as the connected set of 542 proteins linked according to 727 experimentally observed protein–protein interactions from the Database of Interacting Proteins (inset, Figure 4) [3•,35]. These studies reinforce the idea that proteins rarely work in isolation, but are instead linked into an interconnected network of physical interactions and functional relationships.

## Why is this computational genetics?

Unlike sequence homology and inferences from protein structure, nonhomology methods reveal protein function in the same manner that experimental geneticists do: by defining the context that the protein operates in. Function is then determined from the pathway neighbors of a protein. For this reason, we might consider nonhomology methods to be *computational genetics*, a bioinformatics analysis that proceeds in a fashion analogous to experimental genetics.

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

**Figure 4**




The network of 12,012 functional relationships among 2240 proteins from yeast generated from protein phylogenetic profiles, showing links that occur with an expectation (*E*) value $< 1 \times 10^{-3}$. Each vertex represents a protein and each line represents a functional link, modeled as springs to position functionally related proteins close together in space [27••]. In this case, the phylogenetic profiles are calculated for $n = 24$ genomes and the *E* value of a link from protein **A** to protein **B** is calculated as $p(\mathbf{A}) \bullet V(n,d_{\mathbf{AB}}) \bullet N \bullet C$, where $p(\mathbf{A})$ is the probability of observing the profile of protein **A**, $V(n,d_{\mathbf{AB}})$ is the volume of an *n*-dimensional hypersphere centered on **A** of radius $d_{\mathbf{AB}}$, *N* is the number of proteins with informative vectors and *C* is a scale factor. For comparison, the inset shows an experimentally derived network of protein–protein interactions from the Database of Interacting Proteins [35], courtesy of I Xenarios.

In fact, the method of phylogenetic profiles [14•] is an exact computational equivalent of the experimental genetics approach of mapping a mutant gene's phenotype to the gene. When we compare one organism with another, we can generalize each organism as having a collection of mutations, gene knockouts and extra genes relative to the other organisms. By grouping genes with similar phylogenetic profiles, we are mapping genes that produce shared phenotypes (the genes are expressed or absent in the same sets of organisms) and are essentially performing a standard genetic mapping. Of course, the experiment is performed computationally and in a massively parallel form, but it is essentially the same analysis as in experimental genetics.

## Conclusions

This past year has seen an explosion of new experimental and computational tools to identify protein function, including the development of 'nonhomology' computational methods. These methods take advantage of the many properties shared among functionally related proteins, such as patterns of domain fusion, evolutionary co-inheritance, conservation of relative gene position and correlated expression patterns. Such analyses, building on existing genomic sequence and expression data, allow the assignment of preliminary protein function on a genome-wide scale. Even more exciting is the potential for increasing the power of the methods as more genome sequences and expression libraries accumulate; for example, the number of possible phylogenetic profile vectors and, therefore, the potential to separate unrelated proteins grows on the order of $2^n$ for $n$ genomes. An important goal will be working out proper statistical evaluations of results from each of the methods.

In the next year, we can expect these techniques to be integrated, for each genome, with homology- and structure-derived protein functions, as well as with known experimental data, as researchers are beginning to extract experiments from the scientific literature into computer-analyzable databases, such as for protein–protein interactions [35,36], functional relationships derived from the co-occurrence of gene names in articles [37], metabolic pathways [38,39] and general gene function [40–42]. Beyond even these, we can expect many new types of data, such as the recent genome-wide gene disruption phenotypic studies of yeast [43••] and protein expression datasets, that should really open up the power of these methods and allow researchers to finely map many of the functions and relationships among the genes so tantalizingly revealed in each newly sequenced genome.

## Acknowledgements

This work was supported by a Department of Energy/Oak Ridge Institute for Science and Education Hollaender Distinguished Postdoctoral Fellowship and grants from the DOE. The author would like to thank David Eisenberg, Matteo Pellegrini, Michael Thompson, Todd Yeates and Ioannis Xenarios for support and fruitful scientific collaboration.

## References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC *et al.*: **Distinctive gene expression patterns in human mammary epithelial cells and breast cancers.** *Proc Natl Acad Sci USA* 1999, **96**:9212-9217.

2. • Martzen MR, McCraith SM, Spinelli SL, Torres FM, Field S, Grayhack EJ, Phizicky EM: **A biochemical genomics approach for identifying genes by the activity of their products.** *Science* 1999, **286**:1153-1155.
The genome-wide expression and purification of yeast proteins for associating biochemical activities with specific yeast proteins is described.

3. • Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein–protein interactions from genome sequences.** *Science* 1999, **285**:751-753.
This paper shows that domain fusions can be used to predict functionally related and physically interacting proteins.

4. • Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402**:86-90.
The authors developed a scoring scheme for domain fusion analysis that accurately predicts interacting proteins.

5. Park J, Teichmann SA, Hubbard T, Chothia C: **Intermediate sequences increase the detection of homology between sequences.** J Mol Biol 1997, **273**:249-254.

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

DOCKET ALARM

# Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

## Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

## Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

## Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

## API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

**LAW FIRMS**
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

**FINANCIAL INSTITUTIONS**
Litigation and bankruptcy checks for companies and debtors.

**E-DISCOVERY AND LEGAL VENDORS**
Sync your system to PACER to automate legal marketing.