# Increasing the Specificity of Protein Functional Inference by the Rosetta-Stone Method

M. J. Thompson[1], E. M. Marcotte[1,2], M. Pellegrini[1],

T. O. Yeates[1], & D. Eisenberg[1*]

[1] *Protein Pathways, Inc.*

*1145 Gayley Ave., #304*

*Los Angeles, CA 90024*

[2] *Molecular Biology Institute*

*UCLA-DOE Lab of Structural Biology and Molecular Medicine*

*University of California, Los Angeles, P.O. Box 951570*

*Los Angeles, CA 90095-1570*

March 23, 2000

*author to whom correspondence should be addressed

Keywords: Rosetta-Stone, bioinformatics, functional genomics, protein function prediction, Phylogenetic Profiles

# Abstract

The Rosetta-Stone method determines functional interactions among proteins that are not homologous to one another. Given two separate and non-homologous query proteins, the occurence of a third sequence representing the fusion of the two query proteins is taken as evidence of a functional coupling between the two queries. This coupling may represent an actual physical interaction, or it may imply co-participation in the same structural complex or pathway.

The Rosetta-Stone method has shown considerable power in providing functional inferences for a large number of proteins. Problems, however, arise in the automated application of this technique. The main problem stems from the lack of a quantitative measure of the reliability or functional specificity of the links that are generated. This issue stems from the presence of "promiscous domains" that can be found in multi-domain proteins that play roles in diverse cellular processes. A blind application of the Rosetta Stone method will generate vast numbers of non-specific links among these different proteins.

Some method for discriminating between functionally-specific and spurious links is needed. Clearly, sequence alignment scores by themselves do not distinquish promiscuous domains from less common domains, so these cannot be used for this purpose. The other alternative is to identify and cull these domains from the database of interest. Unfortunately this could throw away potentially valuable data as well as require an *ad hoc* definition of what is considered a "promiscuous" domain. [point out that these fusions are unlikely to be garbage, just non-specific]

Here we present a more general approach for addressing this problem. We employ a simple model to compute the probability that any two query proteins would be found to be fused as a third protein, at random. We apply this calculation to all of the Rosetta-Stone linnks generaged among the open reading frames (ORFs) of 26 complete and public genomes. We provide a statistical demonstration that this technique can improve the functional specificity (prediction accuracy) of the Rosetta-Stone method on this large scale. We also provide some examples, including those involving human sequences. [rewrite]

# Introduction

The determination of functions for the vast number of newly sequenced genes and their products has become a problem of central importance for molecular biology in the post-genomic sequence era. The fundamental technique for assigning function to newly sequenced genes and proteins follows a function-by-homology inference [paradigm]. The function of a novel sequence is assumed to be similar to that of sequences to which it is homologous and for which some experimental charactization has been performed. Nearly all traditional bioinformatics techniques, including sequence alignment, motif recognition, hidden Markov Models, and remote homolog detection by fold-recogtion rely on this basic idea. Unfortunately, only a fraction of newly sequenced genes and proteins are homologous to an existing and characterized sequence. This is particularly true when considering the genomes of larger and more complex organisms such as humans. The remaining set of sequences can be classified into those which possess functionally uncharacterized homologs in various organisms ("conserved hypotheticals") and those that apparently have no homologs at all (the "ORFans") [3, 4].

Fortunately, recent innovations in computational protein function assignment have been made that do not rely on the direct function-by-homology inference method. In con-

trast, these techniques infer functional-couplings among non-homologous proteins. Thus, this greatly expands the number of novel sequences for which a function can be supplied. [5, 6, 7, 14, 12, 13].

While the focus of this paper concerns improvements that have been made to the Rosetta-Stone method [6], other advances in computational techniques for sequence functionation that do not rely directly on inference-by-homology have also seen recent and rapid development. The "Phylogenetic Profile" method detects functional-couplings among proteins of a genome by detecting correlated evolution across the spectrum of completely sequenced genomes [5]. The "gene cluster" method looks for evolutionary conservation of gene proximity in the physical genome as an indicator of functional coupling[2, 14]. In addition, all three computational methods have been combined and used with experimental data and literature compilations to assign functions to open reading frames (ORFs) on the genomic level for *Saccharomyces cerevisiae* and *Mycobacterium tuberculosis* [7, 12, 13].

Some problems, however, do arise that complicate the application of the Rosetta-Stone method in an automated manner. The main problem stems from the presence of "promiscous domains" that are incorporated into multi-domain proteins. These domains frequently serve as signalling modules or perform common functions that are necessary in various cellular processes. Multi-domain proteins and promiscuous domains are much more prevalent in the proteomes of higher organisms such as humans and multicellular animals than in microbes. Subsequently, for the genomes of complex organisms, the Rosetta-Stone method generates a combinatorial explosion of functional links.

The links involving these promiscuous domains will be rather non-specific in terms of function. Previous work with the Rosetta-Stone method relied on identifying and culling these domains to alleviate this explosion of non-specific linkages. Here we present a simple probabilistic model that eliminates the need for that drastic measure. Rather than culling

or masking this potentially valuable data, we compute the statistical significance of any link generated by the Rosetta-Stone method.

[hammer Ouzounis method as being too stringent]

We find that this simple model can improve the accuracy of function prediction based on Rosetta-Stone links. A statistical summary of this improved performance is given using the Rosetta-Stone links generated using all the sequences from 26 complete genomes. We demonstrate that it gives appropriately high $p$-values to links involving known promiscous domains, and low $p$-values to links involving proteins whose function is closely connected. We include some human examples. This technique should prove extremely valuable in the functional analysis of genomes from eukarotes, multi-cellular organisms, and humans. [rewrite]

# Methods

For any two query sequences in a sequence database, we would like to compute the probability of randomly observing a third sequence representing the fusion of the two queries. In order to formulate a solution to this problem, it is useful to construct a data structure consisting of "homolog vectors" for the sequences in this database.

## Theory

For each sequence, $s_i$, in a database of $N$ sequences, we construct a binary homolog vector, $\vec{h}_i = (h_{i_1}, \ldots, h_{i_N})$ where the bits $h_{i_l}$ correspond to the $l = 1, \ldots, N$ sequences in the databse. Each bit $h_{i_l}$ denotes the presence (1) or absence (0) of a homology between the query sequence $s_i$ and the database sequence indexed by $l$. Self-homology is neglected by setting $h_{i_i} = 0$ for all $i$. The total number of homologs present in the database for sequence $s_i$ is denoted by $||\vec{h}_i||$ (*i.e.* total number of 1 bits).

If we consider sequences, $s_i$ and $s_j$, that are not homologous to one another ($h_{i_j} = h_{j_i} = 0$) and find, for any other bit position $l$ that $h_{i_l} = h_{j_l} = 1$ then the database sequence $s_l$ corresponds to a fusion of the two sequences. We will denote the number of such "matching 1-bits" as $m_{i,j}$. In this paper, we are interested in the probability of observing a fusion of the two query proteins, not in the probability of observing exactly $m_{i,j}$ fusions. For two sequences, $s_i$ and $s_j$ we can express this as the conditional probability, $P(m_{i,j} > 1 \mid ||\vec{h_i}||, ||\vec{h_j}||)$, of observing at least one matching 1-bit given the respective numbers of homologs (1-bits) $||\vec{h_i}||$ and $||\vec{h_j}||$.

Since the only two possible outcomes are that $s_i$ and $s_j$ have at least one matching homolog or none at all, then the probabilities of these outcomes must sum to 1. We write,

$$P(m_{i,j} = 0 \mid ||\vec{h_i}||, ||\vec{h_j}||) + P(m_{i,j} > 1 \mid ||\vec{h_i}||, ||\vec{h_j}||) = 1 \tag{1}$$

It turns out that from combinatorics considerations, the functional form for the first term on the left-hand side of the equation above is simpler. Thus, we rewrite Equation 1 as

$$P(m_{i,j} > 1 \mid ||\vec{h_i}||, ||\vec{h_j}||) = 1 - P(m_{i,j} = 0 \mid ||\vec{h_i}||, ||\vec{h_j}||) \tag{2}$$

To compute the probability on the right hand side of the equation above, we must first count the total number of ways that the $||\vec{h_i}||$ and $||\vec{h_i}||$ homologs of $s_i$ and $s_j$ can be distributed among the $N$ bit positions of the respective homolog vectors without any matching 1-bits. The product of the two combinatorics functions "$N$ choose $||\vec{h_i}||$" and "$N - ||\vec{h_i}||$ choose $||\vec{h_j}||$" yields this number. In factorial notation, this is given by $\frac{N!}{(||\vec{h_i}||)!(N-||\vec{h_i}||)!} * \frac{(N-||\vec{h_i}||)!}{(||\vec{h_j}||)!(N-||\vec{h_i}||-||\vec{h_j}||)!}$ To obtain a probability, we must then divide by the total number of configurations without placing any restriction on matching 1-bits. This is given by, $\frac{N!}{(||\vec{h_i}||)!(N-||\vec{h_i}||)!} * \frac{N!}{(||\vec{h_j}||)!(N-||\vec{h_j}||)!}$

After cancellation of the "$N$ choose $||\vec{h_i}||$" factorial terms we obtain,

$$P(m_{i,j} = 0 \mid ||\vec{h_i}||, ||\vec{h_j}||) = \frac{(N - ||\vec{h_i}||)!(N - ||\vec{h_j}||)!}{N!(N - ||\vec{h_i}|| - ||\vec{h_j}||)!} \quad (3)$$

Next, we observe that $N$ is typically on the order of tens to hundreds of thousands of sequences. Thus, we can use Stirling's approximation, $log(N!) \approx N(log(N) - 1)$ in computing this probability. We take the log of both sides of the Equation 3, substitute Stirling's approximation, evaluate, re-exponentiate to get the probabilty and substitute into Equation 2.

$$P(m_{i,j} > 1 \mid ||\vec{h_i}||, ||\vec{h_j}||) = 1 - e^{log(P(m_{i,j}=0|||\vec{h_i}||,||\vec{h_j}||))} \quad (4)$$

## Function Keywords

We evaluate the predictive performance of this improved Rosetta-Stone method for detecting functional couplings among proteins using a keyword-recovery statistic developed in the original Rosetta-Stone paper by Marcotte, *et al* [6]. Since the Rosetta-Stone method detects links among non-homologous proteins, we do not generally expect it to provide precise information about the biochemical activity of a query protein. Rather, we expect these links to inform us about the cellular role that the two proteins are playing.

Function keywords that describe the cellular context for the open reading frames of complete genomes can be found at the Kyoto Encyclopedia of Genes and Genomes (KEGG) [9]. These annotations have three levels. The first level describes the major functional category of the ORF (*e.g.* Amino acid metabolism), and the second level describes a minor functional category of the first level (*e.g.* Histidine metabolism). The third level of annotation lists the gene name (if any) and typically the biochemical activity of the sequence (*e.g.* pepD, pepH; aminoacyl-histidine dipeptidase). We obtained these keywords from KEGG for the annotated portions of 24 complete genomes that were available at the time this work was done. As

explained above, keywords denoting gene names or biochemical activity were excluded. [cite Monica Riley?]

Each sequence is then described by two "keywords" being the concatenation of the words describing the major and minor categories, respectively. The keyword recovery between a single pair of linked sequences can either be 0%, 50% or 100%. We also note that some ORFs have multiple dijoint annotations. This may be due to enzymes that catalyze reactions in multiple pathways or uncertainty on the part of KEGG curators as to the specific pathway of the ORF. For links in which one or both ORFs have multiple (or uncertain) functional roles, we compute the keyword recovery between the most similar annotations of the two.

To obtain the keyword recovery, $s_j(q_i)$, for a query ORF, $q_i$, by its $j$th functional partner, $p_{i_j}$, we simply compute, $s_j(q_i) = \frac{k_{i,j}}{\sqrt{k_i * k_j}}$, where $k_i$ and $k_j$ denote the number of keywords for $q_i$ and $p_{i_j}$, respectively, and $k_{i,j}$ denotes the number of keywords shared by the two. To obtain a performance statistic for an entire set of functional links defined by a threshold $p$-value, we average over the $n_i$ functional partners for each query ORF and then average over the $N$ query ORFs,

$$<< s_j(q_i) >_j >_i = \frac{1}{N} \sum_i^N \frac{1}{n_i} \sum_j^{n_i} \frac{k_{i,j}}{\sqrt{k_i * k_j}} \tag{5}$$

## Sequence Data & Alignments

The complete set of open reading frames (ORFs) for 26 microbial and eukaryotic genomes along with the partial set of human sequences were obtained from The National Center for Biotechnology Information (NCBI) [10]. For the analysis of Rosetta-Stone links in the 26 complete genomes, all ORFs from these genomes were aligned against one another using the program PSI-BLAST [1]. The human sequences were run against the BLAST non-redundant database of sequences. Ideally, for any two query proteins, we would want to search the

entire set of available sequences for a Rosetta-Stone protein. For the purposes of testing this approach, however, we have restricted our attention to the comparisons described above.

### Identification of Rosetta-Stone links

From the all vs. all sequence alignment search using the ORFs from the 26 complete genomes, we identified pairs of sequences which had a third common homolog (the putative fusion protein). Since the Rosetta-Stone method aims to identify functional links among non-homologous proteins, we employed stringent filters on these triplets of sequences. We first eliminated those triplets where the two query proteins had some homology to one another. We used a fairly permissive PSI-BLAST $E$-value threshold of $10^{-2}$ for this purpose and did not insist that the homology be detected bi-directionally at this significance level. We then eliminated those triplets where the aligned regions of the pair of proteins against the putative fusion protein overlapped. Finally, only those triplets where each of the pair of proteins aligned bi-directionally with the fusion protein with $E <= 10^{-4}$.

## Results & Discussion

## References


[1] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

[2] Dandekar, T., Snel, B., Huynen, M., & Bork, P., Conservation of gene order: a fingerprint of proteins that physically interact. Trends in Biochemical Sciences **23**, 324-328 (1998).

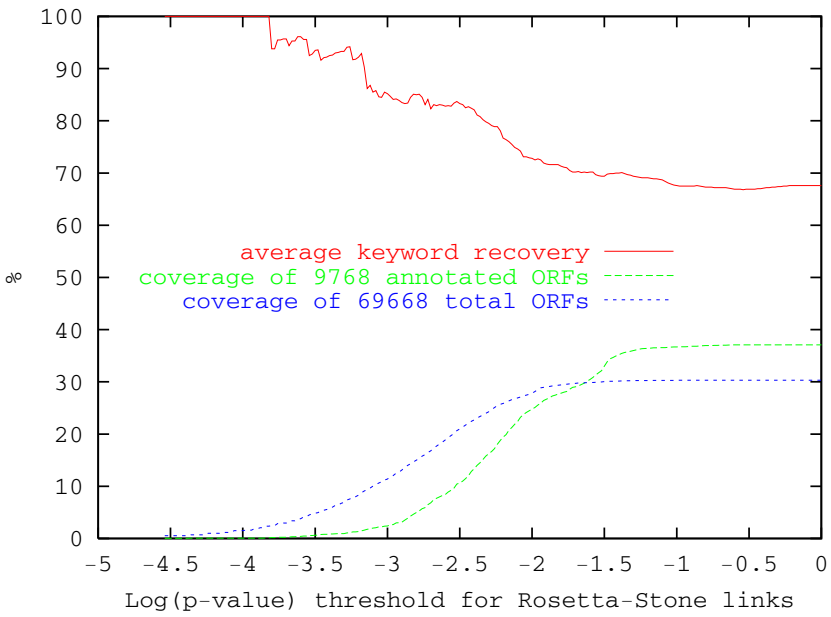


Figure 1: **Statistically significant Rosetta Stone links provide more accurate function prediction.** Rosetta-Stone links were compiled using open reading frames from all 26 complete public genomes. Using a probabilistic model developed at Protein Pathways, a statistical significance was computed for all links. A threshold value of this probability was varied to define subsets of links and the average keyword overlap among annotated pairs in these sets was computed. Note that there has been no previous attempt to score such linkages.

[3] Fischer, D., and Eisenberg D., Finding Families for Genomic ORFans. *Bioinformatics* **15**, 759-762 (1999).

[4] Fischer, D., and Eisenberg D., Structural genomics: Affirmative action for ORFans and the growth in our structural knowledge. *Prot. Eng.* **12**, 101-102 (1999).

[5] 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. Acac. Sci.* **96**, 4285 (1999).

[6] Marcotte, E.M., Pellegrini, M., Ng, H.-L., Rice, D.W., Yeates, T.O., & Eisenberg, D., Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751-753 (1999).

[7] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., & Eisenberg, D., Detecting protein function and protein-protein interactions from genome sequences. *Nature* **402**, 83-86 (1999).

[8] Marcotte, E.M., Xenarios, I., van der Bliek, A. & Eisenberg, D., Discovering organellar proteins from their phylogenetic profiles. *Submitted.* (2000)

[9] `http://www.genome.ad.jp/kegg/`.

[10] `http://www.ncbi.nlm.nih.gov/`.

[11] 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**, 14864-14868 (1998).

[12] Rotstein, S.H., Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., & Eisenberg, D. Computational Assignment of Function to proteins of *Mycobacterium tuberculosis*. *Submitted.* (2000)

[13] Rotstein, S.H., Marcotte, E.M., & Eisenberg, D. Discovering Novel Drug Targets in Mycobacterium tuberculosis *Mycobacterium tuberculosis. Submitted.* (2000)

[14] Overbeek, R., Fonstein, M., D'Souze, M., Pusch, G.D., & Maltsev, N. The use of gene clusters to infer functional couplings. *Proc. Natl. Acad. Sci. USA* **96**, 4285-4288 (1999).

[15] Ewing, R.M., Kahla, A.B., Poirot, O., Lopez, F., Audic, S., & Claverie, J-M. Large-Scale Statistical Analyses of Rice ESTs Reveal Correlated Patterns of Gene Expression. *Genome Research* **9**, 950-950.

[16] Walker, M.G., Volkmuth, W., and Klingler, T.M., Pharmaceutical target discovery using Guild-by-Association: schizophrenia and Parkinson's disease genes.

[17] Walker, M.G., Volkmuth, W., Sprinzak, E., Hodgson, D., Klinger, T., Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res* **12**, 1198-1203 (1999).

[18] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J.M., A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623-631.

[19]