

Data fusion for disease gene identification (and a brief primer about gene networks)

Victor Hanson-Smith
CIS607 Winter 2009

The central dogma of biology. . .

Some regions of DNA are *coding-regions*, which encode for a gene.

DNA strand: ATCBATTABCATCG. . . TTGCTAGCTATTTACAGCATCACATCTAATCGATCGCTTCTAGC

A type of protein called *transcriptase* transcribes coding-regions of DNA into RNA.

RNA: AUGC. . . UUGC

The ribosome (which is a polymer of proteins) translates messenger RNA (mRNA) into proteins.



Protein molecules are diverse!

The human body is estimated to contain 25,000 different types of proteins. Here are a few of my favorite proteins, which don't necessarily appear in the human body:

Glycoprotein antifreeze protein in ice fish

<http://www.pnas.org/content/94/8/3811.abstract>

Steroid-hormone receptor protein in vertebrates:

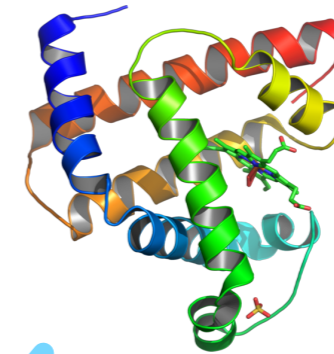
<http://www.sciencemag.org/cgi/content/abstract/312/5770/97>

Opsin visual pigment protein in archosaurs:

<http://mbe.oxfordjournals.org/cgi/content/abstract/19/9/1483>

Some types of protein regulate the transcription & translation of other proteins. . .

Some regions of DNA are *promoter regions*, *enhancer regions*, or *silencer regions*.



DNA strand: ATC.....TABCATCG. . . TTGCTAGCTATT.....GCATCAGATCTAATCGATCGCTTCTAGC

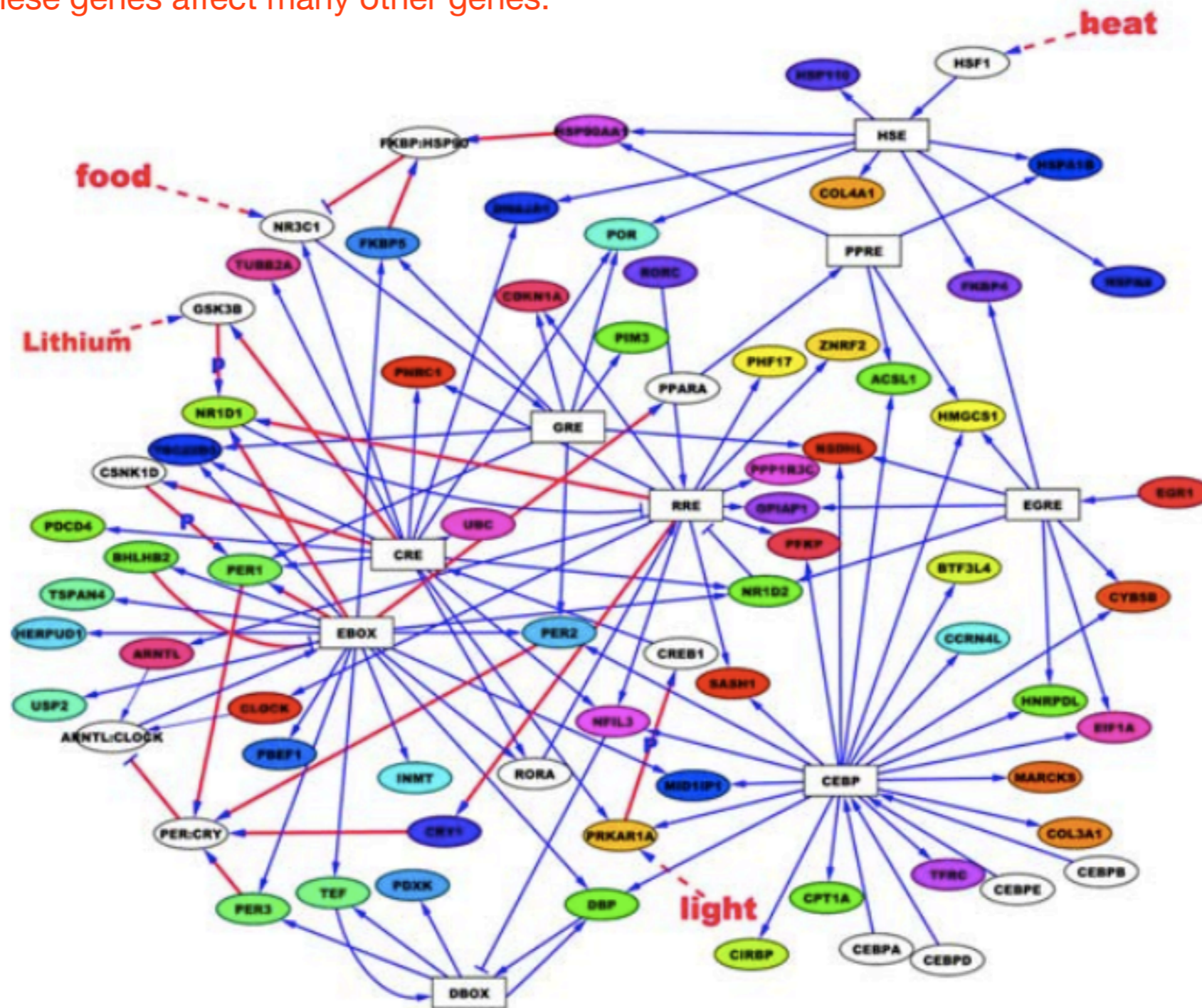


When certain types of protein bind to promoter or enhancer regions, the related coding-region is able to be transcribed.

When proteins bind to silencer regions, the related coding-region will not be transcribed.

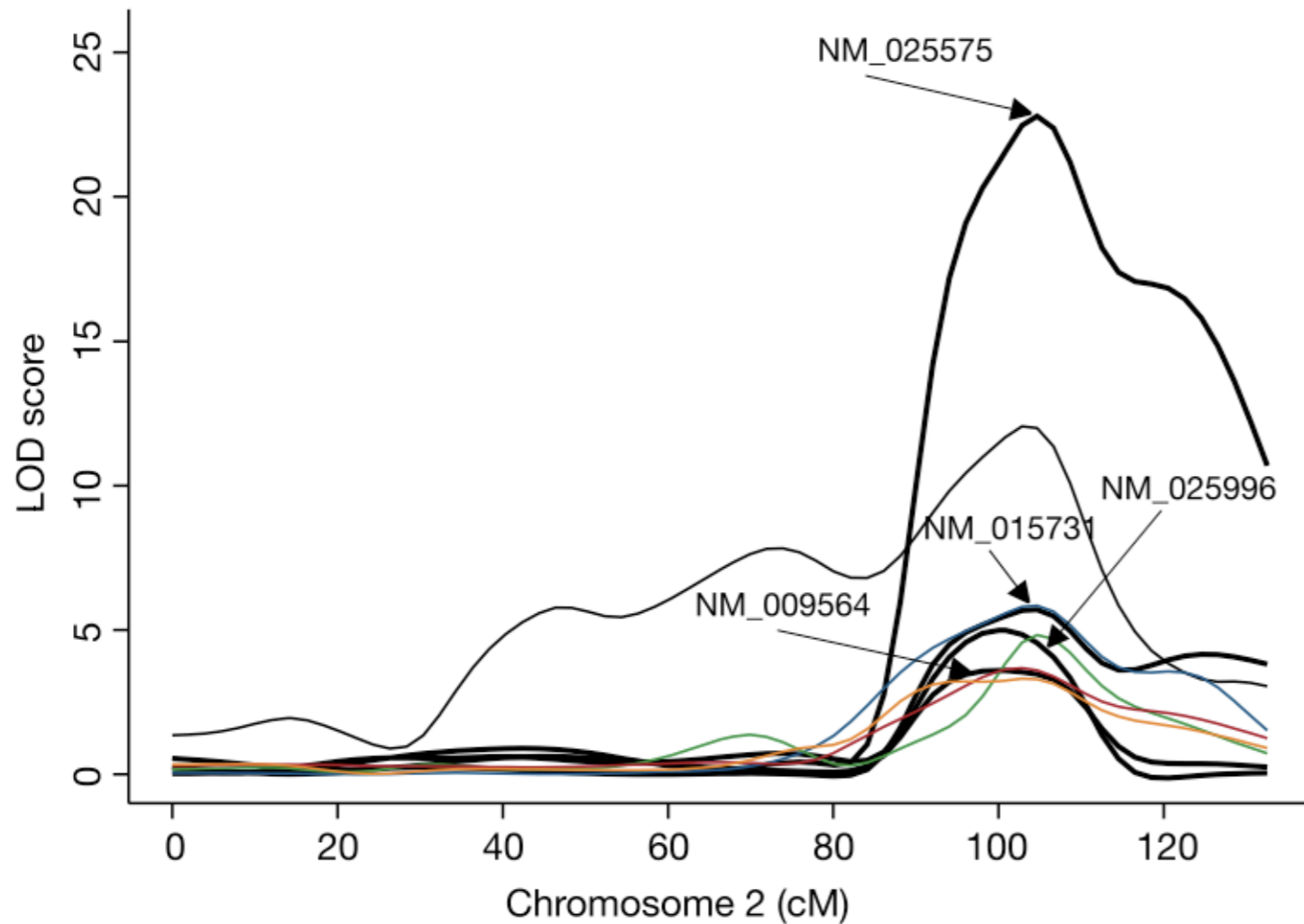
The big picture: proteins interact in a network

The downstream phenotype is determined by a complex interaction of proteins promoting, enhancing, and/or silencing the expression of other proteins. For example, consider the circadian rhythm in mammals (below): heat, food, and light directly affect only a few genes, but these genes affect many other genes.



Microarray analysis measures the expression levels of RNA

By comparing the different RNA expression levels between populations, we can potentially identify which genes are responsible for the differences between the populations.



Here, researchers compare the RNA expression levels between obese and non-obese populations. They identify a set of genes which might be responsible for obesity.

These putative genes are called quantitative trait loci, or **QTL**.

One challenge with QTL/microarray analysis is identifying the root gene which causes a disease

Although QTL identifies several genes which might be the cause of some condition (i.e. a disease), some of these genes might actually be downstream in the gene-interaction network.

This week's paper. . .

Gene prioritization through genomic data fusion

Stein Aerts^{1,4,5}, Diether Lambrechts^{2,5}, Sunit Maity^{2,5}, Peter Van Loo³⁻⁵, Bert Coessens^{4,5}, Frederik De Smet², Leon-Charles Tranchevent⁴, Bart De Moor⁴, Peter Marynen³, Bassem Hassan¹, Peter Carmeliet² & Yves Moreau⁴

Aerts et al., 2006, Nature Biotechnology

By using data fusion, Aerts et al. are able to better identify which genes cause a disease versus the genes which are simply related in the network.

Their approach. . .

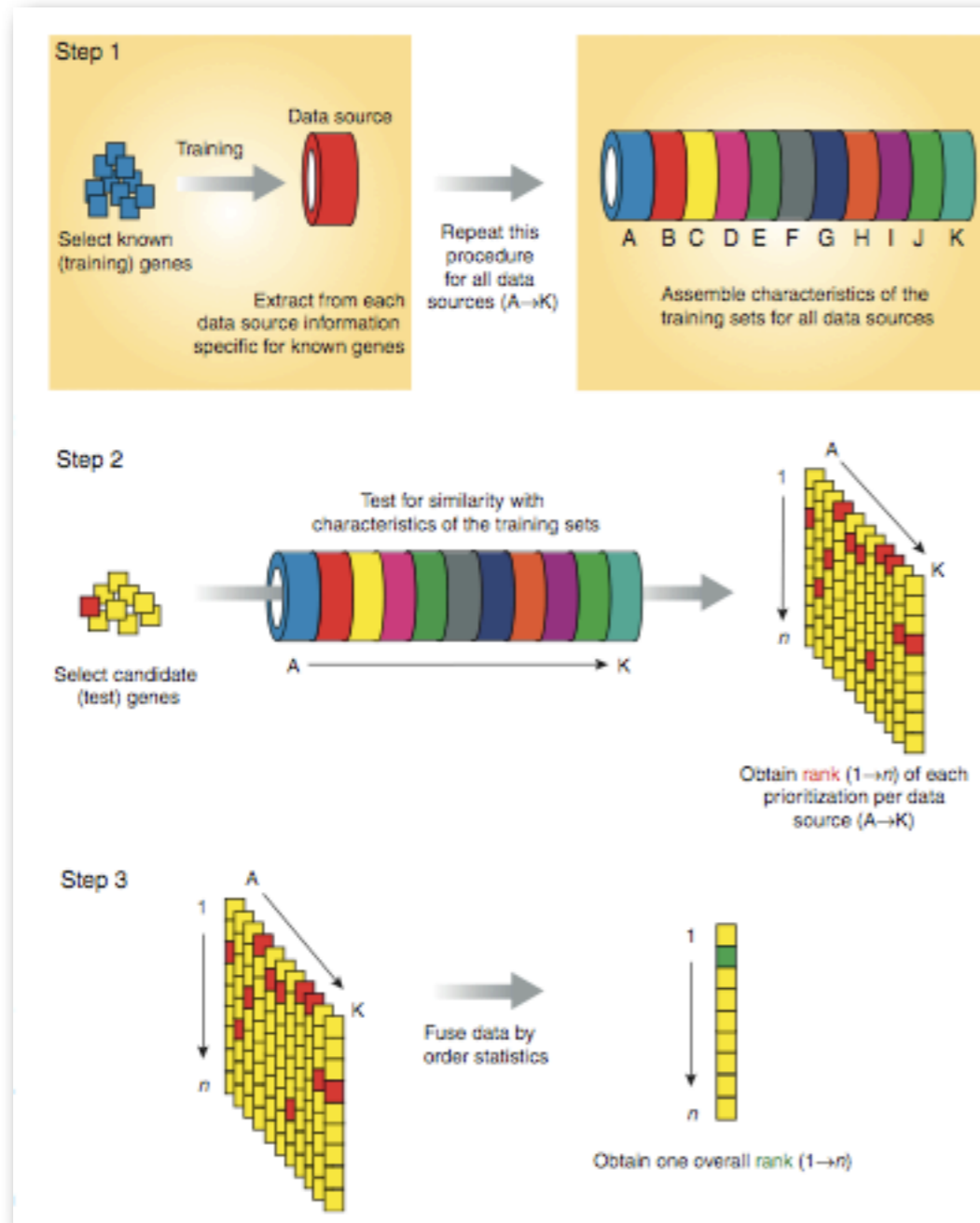


Figure 1 Concept of prioritization by Endeavour. Candidate test genes are ranked with Endeavour based on their similarity with a set of known training genes in a three-step analysis. In the first step (upper panel), information about a disease or pathway is gathered from a set of known training genes by consulting various data sources. Training genes can be loaded automatically (based on a Gene Ontology term, a KEGG pathway ID or an OMIM disease name) or manually. The latter allows the incorporation of expert knowledge. The following data sources are used: A, literature (abstracts in EntrezGene); B, functional annotation (Gene Ontology); C, microarray expression (Atlas gene expression); D, EST expression (EST data from Ensembl); E, protein domains (InterPro); F, protein-protein interactions (Biomolecular Interaction Network Database or BIND); G, pathway membership (Kyoto Encyclopedia of Genes and Genomes or KEGG); H, *cis*-regulatory modules (TOUCAN); I, transcriptional motifs (TRANSFAC); J, sequence similarity (BLAST); K, additional data sources, which can be added (e.g., disease probabilities). In the second step (middle panel), a set of test genes is loaded (again, either manually or automatically by querying for a chromosomal region or for markers). These test genes are then ranked based on their similarity with the training properties obtained in the first step, which results in one prioritized list for each data source. Vector-based data are scored by the Pearson correlation between a test profile and the training average, whereas attribute-based data are scored by a Fisher's omnibus analysis on statistically overrepresented training attributes. Finally, in the third step (lower panel), Endeavour fuses each of these rankings from the separate data sources into a single ranking and provides an overall prioritization for each test gene. As such, Endeavour prioritizes genes through genomic data fusion—a term, borrowed from engineering to reflect the merging of distinct heterogeneous data sources, even when they differ in their conceptual, contextual and typographical representations.

See also. . .

BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i125–i132
doi:10.1093/bioinformatics/btm187

Kernel-based data fusion for gene prioritization

Tijl De Bie^{1,2,*}, Léon-Charles Tranchevent³, Liesbeth M. M. van Oeffelen³ and Yves Moreau³

¹Department of Engineering Mathematics, University of Bristol, University Walk, BS8 1TR, Bristol, UK, ²OKP Research Group, Katholieke Universiteit Leuven, Tiensestraat 102, 3000 Leuven, Belgium and ³ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

Published online 28 May 2008

Nucleic Acids Research, 2008, Vol. 36, Web Server issue W377–W384
doi:10.1093/nar/gkn325

ENDEAVOUR update: a web resource for gene prioritization in multiple species

Léon-Charles Tranchevent¹, Roland Barriot¹, Shi Yu¹, Steven Van Vooren¹, Peter Van Loo^{1,2,3}, Bert Coessens¹, Bart De Moor¹, Stein Aerts^{3,4} and Yves Moreau^{1,*}

¹Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, ²Human Genome Laboratory, Department of Molecular and Developmental Genetics, VIB, Leuven, ³Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine and ⁴Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, Leuven (Belgium)

Received February 7, 2008; Revised April 30, 2008; Accepted May 7, 2008

Int. J. Biol. Sci. 2007, 3

420

International Journal of Biological Sciences

ISSN 1449-2288 www.biolsci.org 2007 3(7):420-427

©Ivyspring International Publisher. All rights reserved

Review

Candidate Gene Identification Approach: Progress and Challenges

Mengjin Zhu and Shuhong Zhao

Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, Huazhong Agricultural University, Wuhan 430070, P. R. China

fin.