Article III

# Context dependent visualization of protein function

In: Juho Rousu, Samuel Kaski and Esko Ukkonen (eds.). Probabilistic Modeling and Machine Learning in Structural and Systems Biology. 2006, Tuusula, Finland, pp. 26–31.

# Context dependent visualization of protein function

Peddinti V. Gopalacharyulu[1], Erno Lindfors[1], Catherine Bounsaythip[1], and Matej Orešič[1]

[1] VTT Technical Research Centre of Finland, Tietotie 2, FIN-02044 Espoo, Finland
{ext-peddinti.gopal, erno.lindfors, catherine.bounsaythip, matej.oresic}@vtt.fi

**Abstract.** Assignment of protein function is a nontrivial task due to the fact that the same proteins may be involved in different biological processes, depending on the state of the biological system and protein localization. Therefore, protein function is context dependent and textual annotations commonly utilized to describe protein function lack the flexibility to address such contextuality. We propose an alternative approach for protein annotation motivated by the conceptual space approach, which relies on context-driven mapping of complex relationships based on known protein interactions or ontologies and on experimental data into low-dimensional space. We utilize the curvilinear distance analysis to generate such mappings, and demonstrate the approach on a set of proteins involved in maintenance of energy homeostasis.

**Keywords:** Protein function, conceptual spaces, curvilinear distance analysis

## 1 Introduction

The wealth of information generated with modern life science technologies, combined with existing repositories of knowledge dispersed across numerous databases and literature, demand new solutions for management and integration of life science data. Biological systems are characterized by the complexity of interactions of their internal parts and also with the external environment. The protein repositories such as UniProt [1] describe the protein function in textual format. In such form it may be difficult if not impossible to express the protein function in a given context, therefore another layer of representation is necessary.

While biological ontologies such as Gene Ontology [2] attempt to unify part of our life science knowledge at the molecular level, the diversity of life science research and questions addressed inevitably lead to multiple and overlapping ontologies. In turn, these Ontologies need to be integrated and unified, a challenge addressed by the Semantic Web approaches. However, these approaches are mostly based on hard coded symbolic representations which are valid only if the context in which they were created is stable. Therefore, in the fast evolving knowledge in life science, such approaches lack flexibility, emergence and context sensitivity.

In this paper we propose a visual approach for context-dependent protein function characterization, motivated by P. Gärdenfors' paradigm of *conceptual spaces* [3]. The

main idea behind conceptual spaces is that if we use a group of objects or "*clusters*" as references, they are much more reliable than single objects. In the conceptual spaces, *clusters* remain stable even when objects change their properties or when new objects come into existence or old ones disappear. Unlike in ontological structures, the name that is given to a cluster does not need to be taken as such by its sole semantic sense, but it is enriched by the set of qualities (called "*quality dimensions*") of the cluster it represents. Therefore, naming convention is not a bottleneck as in Semantic Web approach.

In living systems the quality dimensions may correspond to different levels of biological organization, where the objects (e.g. molecules, cells, organs) and their quality dimension specific relationships can be described with certain geometric structures (in some cases they are *topological* or *orderings*). Therefore, with the aid of the dimensions, similarities between biological entities and concepts can easily be represented by the distance in a conceptual space.

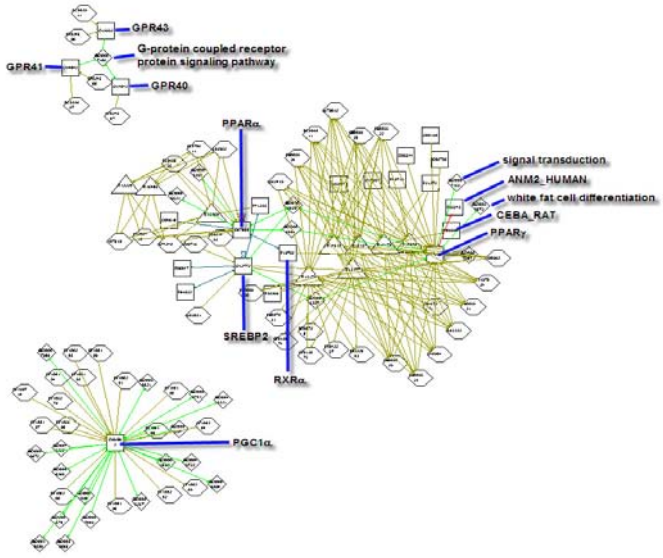


**Fig. 1.** Query for the protein neighborhood of PPARγ, PPARα, PGC1α, SREBP2, GPR40, GPR41, GPR43 human proteins, utilizing BIND [4], MINT [5], DIP [6], KEGG [7], Transfac [8] and Gene Onology [2] databases. Squares represent proteins, hexagons genes, triangles DNA binding sites, and diamonds GO terms.

## 2 Network representation

We represent networks as directed weighted graphs where biological entities are nodes connected via interactions or relationships between them [9]. In the context of protein function, a typical question utilizing network representation is about the *protein neighborhood*, i.e. what are the nearest nodes connected to a particular

protein, with the edges being either direct interactions or ontology-defined relationships. Fig. 1 shows an example of a specific query for a set of human proteins related to maintenance of energy homeostasis and specific G-protein coupled receptors (GPCRs) that are not yet well characterized. The query for the nearest neighbor protein relationships revealed three distinct clusters, with all three GPCRs jointly in a separate cluster. While some of the well known relationships were revealed in the largest cluster, the results of the query have not facilitated characterization of poorly annotated proteins such as GPR40, GPR41, and GPR43.

## 3 Conceptual spaces and biological entities

In the network view illustrated above, a logical follow-up query would include extension of protein neighborhood search for the next-nearest neighbors or beyond. However, due to high-connectivity of biological entities such approach soon becomes visually prohibitive. As a pragmatic alternative, we define a distance metric for each type of relationship, and allow the user to assign weights to different types of relationships as part of the mining process [9]. The key problem then becomes how to efficiently map data to a lower dimensional space in order to be able to visualize them in a context-dependent manner. We implemented Curvilinear Distance Analysis (CDA) [10] in our system. Curvilinear distance depends not only on the two points between which the distance is measured but also on the other surrounding points. Intuitively, instead of computing straight distances between the points, the goal of curvilinear distance consists in computing distances along an object that can be, for example, curves on the surface or any set of points.

CDA maps the points in a higher dimensional space into a lower dimensional space by preserving the distances in the original space. It calculates curvilinear distances in the high dimensional input space by creating a graph out of centroids. After that it calculates distances between centroids using Dijkstra's shortest path algorithm [11]. CDA works by optimizing a criterion that explicitly measures the preservation of the pairwise distances:

$$E_{CDA} = \sum \left(\delta_{ij} - d_{ij}\right)^2 F(d_{ij}, \lambda),$$

where $\delta_{ij}$ is distance measured between points $p_i$ and $p_j$ in the high dimensional data space and $d_{ij}$ is distance measured between the coordinates of the same two points in the projection space. The factor $F(d_{ij},\lambda)$ weighs the contribution of each pair of points in the criterion. $F$ is implemented as the Heaviside unit step function:

$$F\left(d_{ij}, \lambda\right) = \theta\left(\lambda - d_{ij}\right) = 0 \text{ if } \lambda - d_{ij} < 0$$
$$= 1 \text{ if } \lambda - d_{ij} \geq 0 .$$

Starting from the criterion, the derivation of the learning rule follows a similar scheme as for a stochastic gradient descent. Instead of moving one mapped point

according to the position of all other ones, one point $m_i$ is frozen while moving all others radially around it:

$$m_j \leftarrow m_j + \alpha F\left(d_{ij}, \lambda\right)\left(\delta_{ij} - d_{ij}\right)\frac{m_j - m_i}{d_{ij}} \ ,$$

where $\alpha$ and $\lambda$ are time decreasing learning rate and neighborhood radius, respectively.

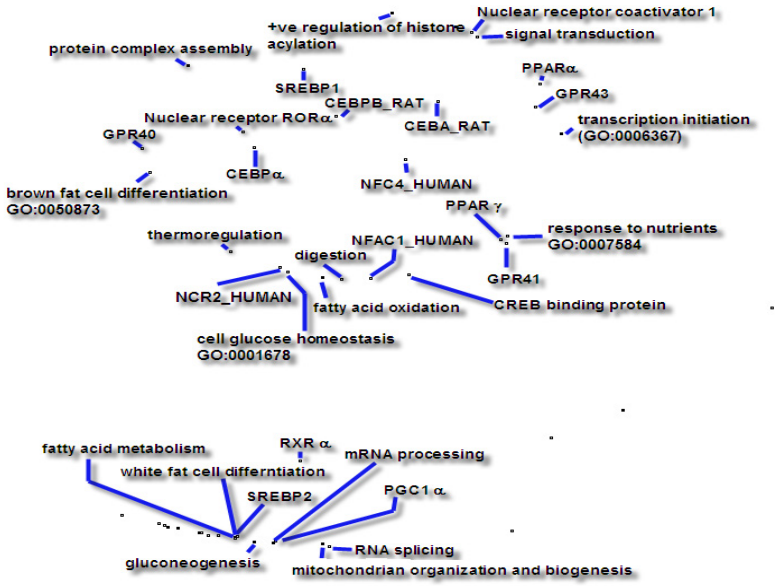


**Fig. 2.** Results of CDA mapping for the entities and databases listed in Fig. 1. All edge weights (unit costs) taken to be 1.

An application example of CDA mapping is shown in Fig. 2, querying for the same entities as in Fig. 1. While there are many interesting aspects of biology retrieved in the mapping, we focus here on PPARγ. PPARγ (UniProt id: P37231) is annotated in UniProt as "*Receptor that binds peroxisome proliferators such as hypolipidemic drugs and fatty acids. Once activated by a ligand, the receptor binds to a promoter element in the gene for acyl-CoA oxidase and activates its transcription. It therefore controls the peroxisomal beta-oxidation pathway of fatty acids. Key regulator of adipocyte differentiation and glucose homeostasis*". This is not a satisfactory explanation when searching for specific context, for example specific disease or relationship to specific GPCR. Our CDA projection revealed both PPARγ and GPR41 are closely associated with response to nutrients. Interestingly, this finding is supported by recent research [12], yet it cannot be revealed by searching any of the

databases used individually.

Our approach is not limited only to pathway data and ontologies. Experimental data, such as gene expression or metabolomics experiments, can also be utilized to further define the context. In such cases the distance measure relating biological entities in the molecular profile space may correspond to the measure of co-expression (such as correlation coefficient) between different entities. Fig. 3 shows an example of CDA mapping based on a similar query as listed previously, but for the mouse proteins, and in the context of a specific gene expression dataset [13] from spleen tissue of NOD mouse. Curiously, several tumor suppressor genes such as BRCA1 associated with PPARγ, are found in this mapping. This finding deserves further attention. Only recently a link between a specific tumor suppressor (LKB1) and diabetes has been established [14], linking cancer and physiological control of metabolism.

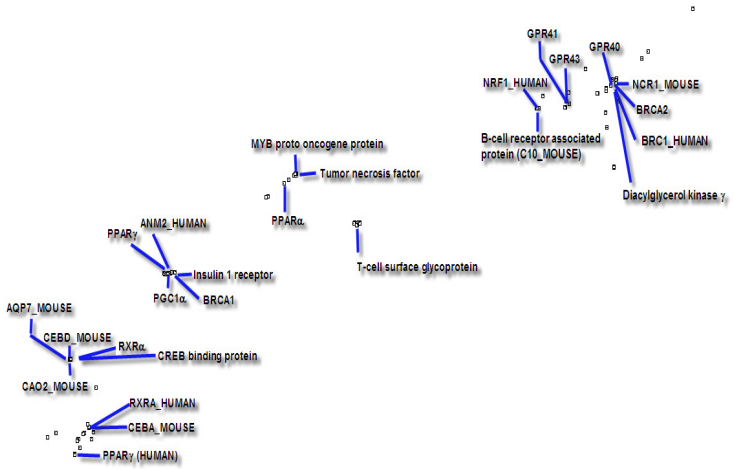


**Fig. 3.** Results of CDA mapping in context of Type 1 Diabetes for mouse proteins PPARγ, PPARα, PGC1α, GPR40, GPR41, GPR43 from databases listed above, plus the gene expression data from [13].

## 4 Conclusions

In this paper we introduced an approach aiming to facilitate mining of complex biological networks, ontologies, and high-dimensional molecular profile data. We

focused specifically on context-dependent protein function assignment. The approach relies on network-based representation of biological entities, concepts, and their relationships, context-dependent assignment of distances between them, and nonlinear mapping into low-dimensional space to visualize distribution of concepts and entities in a specific context. Given the complexity of biological systems and fragmentation of biological knowledge, we believe our pragmatic approach is superior to more formal approaches such as based on Semantic Web technology in its flexibility and ability to extract potentially novel biological relationships leading to new hypotheses. For example, none of the surprising context-dependent functional relationships related to the PPARγ protein shown in this paper could be derived by mining Gene Ontology or other bioinformatics databases alone. Our approach also provides new opportunities for research of topological structures defined by complex biological relationships.

## References


1. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Research, 2004. **32**(Database issue): p. D115-119.
2. Ashburner, M., et al., *Gene ontology: tool for the unification of biology.* Nat. Genet., 2000. **25**: p. 25 - 29.
3. Gärdenfors, P., *Conceptual spaces: The geometry of thought.* 2000, Cambridge, MA: MIT Press.
4. Bader, G.D., D. Betel, and C.W.V. Hogue, *BIND: the Biomolecular Interaction Network Database.* Nucl. Acids Res., 2003. **31**(1): p. 248-250.
5. Zanzoni, A., et al., *MINT: a Molecular INTeraction database.* FEBS Lett., 2002. **513**(1): p. 135-140.
6. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update.* Nucl. Acids Res., 2004. **32**(suppl_1): p. D449-451.
7. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucl. Acids Res., 2004. **32**(90001): p. D277-280.
8. Matys, V., et al., *TRANSFAC: transcriptional regulation, from patterns to profiles.* Nucl. Acids Res., 2003. **31**(1): p. 374-378.
9. Gopalacharyulu, P.V., et al., *Data integration and visualization system for enabling conceptual biology.* Bioinformatics, 2005. **21**(suppl_1): p. i177-185.
10. Lee, J.A., A. Lendasse, and M. Verleysen, *Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis.* Neurocomputing, 2004. **57**: p. 49-76.
11. Dijkstra, E., *A note on two problems in connexion with graphs.* Numerische Mathematik, 1959. **1**: p. 269-271.
12. Xiong, Y., et al., *Short-chain fatty acids stimulate leptin production in adipocytes through the G protein-coupled receptor GPR41.* PNAS, 2004. **101**(4): p. 1045-1050.
13. Eaves, I.A., et al., *Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: The NOD model of Type 1 Diabetes.* Genome Research, 2002(12): p. 232-243.
14. Shaw, R.J., et al., *The Kinase LKB1 mediates glucose homeostasis in liver and therapeutic effects of metformin.* Science, 2005. **310**(5754): p. 1642-1646.