Methods xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

A guide for building biological pathways along with two case studies: hair and breast development

Daniel Trindade [a,1], Lissur A. Orsine [a,1], Adriano Barbosa-Silva [b], Elisa R. Donnard [c], J. Miguel Ortega [a,*]

[a] Depto. de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG 31270-010, Brazil
[b] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette 4362, Luxembourg
[c] Centro de Oncologia Molecular, Instituto de Ensino e Pesquisa, Hospital Sírio-Libanês, São Paulo, SP 01308-060, Brazil

ARTICLE INFO

Article history:
Received 5 May 2014
Received in revised form 26 August 2014
Accepted 3 October 2014
Available online xxxx

Keywords:
Pathway
PubMed
PESCADOR
Hair development
Breast development

ABSTRACT

Genomic information is being underlined in the format of biological pathways. Building these biological pathways is an ongoing demand and benefits from methods for extracting information from biomedical literature with the aid of text-mining tools. Here we hopefully guide you in the attempt of building a customized pathway or chart representation of a system. Our manual is based on a group of software designed to look at biointeractions in a set of abstracts retrieved from PubMed. However, they aim to support the work of someone with biological background, who does not need to be an expert on the subject and will play the role of manual curator while designing the representation of the system, the pathway. We therefore illustrate with two challenging case studies: hair and breast development. They were chosen for focusing on recent acquisitions of human evolution. We produced sub-pathways for each study, representing different phases of development. Differently from most charts present in current databases, we present detailed descriptions, which will additionally guide PESCADOR users along the process. The implementation as a web interface makes PESCADOR a unique tool for guiding the user along the biointeractions, which will constitute a novel pathway.

1. Introduction

Most of the genomic information studied to date is underlined in the format of biological pathways [1]. In its essence, a typical biological pathway to be promptly remembered in biochemistry is the energy metabolism glycolysis pathway. A personal communication of Minoru Kanehisa, founder of KEGG [2] (www.genome.jp/kegg), in a pizza restaurant in São Paulo, revealed the starting motivation for this database: to represent *Escherichia coli* glycolysis and other metabolic pathways from other microbes that were being sequenced, in the year of 1995 [3]. As most readers will be aware, the KEGG database kept on creating pathway-like charts and today, an inspection of these provides great help with the interpretation of some phenomena of interest to the user. For instance, people interested in comparative genomics can, even before they start sequencing their favorite genome, begin inspecting, let's say, DNA repair mechanisms (KEGG pathway maps: map03410, map03420 and map03430) which will be either present or absent in a species related to the organism of choice. Other people perhaps interested in comparing gene expression in two situations, can determine which pathways are enriched or moderated after the treatment. Actually, the current interpretation for KEGG and other pathway databases is the representation of systems, therefore contributing for the study of systems biology.

What are the non-usual pathways already depicted in pathway databases? Reactome (named after a collective for reactions, like genome, proteome, etc.) is a database of pathways [4] authored by expert biologists, following their interests and expertise, standardized and integrated with other databases in collaboration with the Reactome editorial staff. Their homepage states: "The rationale behind Reactome is to convey the rich information in the visual representations of biological pathways familiar from textbooks and articles in a detailed, computationally accessible format". In Reactome one can find charts such as: (i) Circadian Clock, comprising 99 proteins; (ii) Reproduction–Fertilization, with 28 proteins; (iii) under the disease category, HIV infection, subdivided in the

Abbreviations: PMID, PubMed ID; LCA, last common ancestor; HF, hair follicle; MX, matrix; MD, medulla; ORS, outer root sheath; IRS, inner root sheath; DP, dermal papilla; CX, cortex; CL, cuticle; EMT, epithelial to mesenchymal transition.

* Corresponding author.
E-mail addresses: daniel-trindade@c-bio.grad.ufmg.br (D. Trindade), lissurorsine@gmail.com (L.A. Orsine), adriano.barbosa@uni.lu (A. Barbosa-Silva), edonnard@mochsl.org.br (E.R. Donnard), miguel@icb.ufmg.br (J.M. Ortega).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ymeth.2014.10.006
1046-2023/

HIV life cycle, with 340 proteins, and host interactions of HIV factors, with 170 proteins. Therefore, the trend that started out with *E. coli* glycolysis back in 1995 is moving further.

A third resource deserving of mention is WikiPathways [5], established to facilitate the contribution and maintenance of pathway information by the biology community. In there one can find charts on human processes including spinal cord injury, codeine and morphine metabolism, and cardiac progenitor differentiation, with respectively 118, 9 and 53 gene products.

Thus, building biological pathways is an ongoing demand and might largely benefit from a method for extracting information from biomedical literature with the aid of text-mining tools [6]. For the purpose of designing biological pathways, several computational representation models have been developed: (i) markup languages such as KGML (www.kegg.jp/kegg/xml/) used in KEEG database; (ii) universal models such as Biological Expression Language (BEL, www.openbel.org); (iii) or even completely dedicated and inter-operational languages as is the case of Systems Biology Markup Language (SBML, www.sbml.org). The biological information represented using these models might be loaded to downstream applications for visualization, edition or expansion. For that purpose, software such as Cytoscape [7], CellDesigner [8] or PathVisio [9] are broadly used by the pathways design community.

Here we hopefully guide you in the attempt of building a customized pathway or chart representation of a system. Our manual is based on a group of software designed to look at biointeractions in a set of abstracts extracted from PubMed. However, they aim to support the work of someone with good biological background, who will play the role of manual curator while designing the representation of the system. The interesting characteristic of this approach is that the curator does not need to be an expert on the subject, but otherwise interested in depicting the phenomenon for further studies. We therefore illustrate with two cases. They were chosen to focus on recent acquisitions of human evolution. If one has information on the distribution of orthologues of the listed genes, it is possible to depict the acquisition of genes along evolution that compose the system or process of interest [10,11].

This manual will guide you along with the necessary care to address finding a useful PubMed query, inspecting a comprehensive and not so large set of abstracts and at last, mining gene names associated by biointeraction terms. Further than obtaining an interaction network, this manual aims to teach how to attain a chart for the pathway of desire. Additionally, we will comment on some approaches that might lead to the definition of the evolutionary history of such a pathway.

The approach suggested here has been undertaken in previous work. Our first case study was instigated by a desire to build a KEGG-like pathway for the preimplantation embryo development [10]. We could summarize what happens in the embryo Inner Cell Mass (ICM), where the phosphorylation of Yap transcription co-activator releases it from the association with Tead4, thus avoiding the induction of the Cdx2 kinase. This same kinase is expressed in the trophectoderm, and represses Nanog and Oct4, therefore releasing the action of genes implicated in the differentiation of this follicle. Remarkably, Nanog is more ancient, with homologues shared with all Coelomata organisms including *Drosophila melanogaster*, while Oct4 humans share just with fish and other Euteleostomi organisms, so it is a recent acquisition of this system [10]. Another previous case was included in the publication of PESCADOR software [12], used in this tutorial. In it, an existing KEGG pathway, Colorectal Cancer, was roughly doubled with additional genes, therefore suggesting that users interested on already published pathways might place some effort on supplementing it.

Here we give all tips that we use in our procedure and present two cases representing challenging subjects, phenomena restricted to recent epochs of human evolution, which are not covered by charts in the available literature.

2. Methods

The initial step for text-mining pathway construction is selecting a set of 3–10 relevant references on the subject of interest. These references should contain highly relevant gene regulation descriptions and will serve as the basis for the subsequent query. For the Hair Development pathway, the initial list consisted of 10 references (PMIDs: 12533516, 11751679, 14962104, 20944652, 15630473, 22506014, 16481354, 15245425, 23602386, 10767081) and for the breast development pathway we selected seven references (PMIDs: 15351091, 17877612, 16807800, 16861925, 18947364, 15351091 and 15886886). Considering that the curator may not be an expert on the subject, this step might be repeated as he/she faces seminal abstracts along the project, including them and repeating the procedure. The best approach for a non-expert in selecting these reference articles is searching for them in the citations of the major reviews in the field. The major reviews will most likely encompass the main interactions established by different research groups, which will form the basis for the abstract ranking step (see below).

The next step is extracting the relevant literature from the PubMed database (www.ncbi.nlm.nih.gov/pubmed). The query should be carefully selected and exclude possibly misleading results that will hinder the text mining approach. The queries used for the two pathways described here were "(hair AND follicle) NOT (ear OR auditory)" and "(breast AND development) NOT cancer". Another suggestion is to start with only few "seed" articles and employing NCBI PubMed's existing "Related Articles" search functionality. The list of PubMed IDs (PMIDs) resulting from the PubMed search should be retrieved as a text file. We obtained 6452 and over 40000 PMIDs for our two queries, respectively.

An efficient and appropriated (i) choice of relevant references and (ii) PubMed query as well is expected to produce many abstracts with marked genes in the final step, while using PESCADOR (below). If this does not occur, the user can adjust both procedures.

The results from the literature search can be overwhelming and difficult to prioritize. The next step makes use of the Medline Ranker software (cbdm.mdc-berlin.de/tools/medlineranker/) to classify the full set of references based on their relevant information content. At this point the initial list of 3–10 references is used as input for the training set. The background or test set is the full PMID list retrieved from your PubMed query. Medline Ranker was used to classify the results from both queries and 1000 references with higher scores (p-value < 0.01) were selected as input to the second software PESCADOR (cbdm.mdc-berlin.de/~pescador/). PESCADOR adds web interface and several features on top of LAITOR software [13] designed to depict biointeractions between terms (genes).

The engine underneath PESCADOR, LAITOR, was designed to be used as command line software, and it depicts co-occurrence analysis of terms (either genes or concepts) with a dictionary of biointeractions (such as induces, represses, regulates, etc.). Co-occurrences are classified into four types: (i) type 1, when the structure is term 1 – biointeraction – term 2, in a single sentence; (ii) type 2, when both terms are in the same sentence but the biointeraction is not between them; (iii) type 3, when both terms are in the same sentence, but no biointeraction is found in the internal dictionary of biointeractions and (iv) type 4, when two terms are found in the abstract, in different sentences.

The PESCADOR online platform allows the user to input not only the list of PMIDs but also customized biological concepts that