Universidade de Lisboa
Faculdade de Ciências
Departamento de Informática

Cell biology informatics

Two bioinformatic tools for the study of evolutionary cell biology

por

Filipe Bernardes da Silva Tavares Cadete

Mestrado em Tecnologias de Informação aplicadas às Ciências Biológicas e Médicas

2009

Universidade de Lisboa
Faculdade de Ciências
Departamento de Informática

Cell biology informatics

Two bioinformatic tools for the study of evolutionary cell biology

projecto realizado no

Instituto Gulbenkian de Ciência

por

Filipe Bernardes da Silva Tavares Cadete

Projecto orientado pelo Prof. Dr. André Osório e Cruz de Azerêdo Falcão
e co-orientado pelo Dr. José Bártholo Pereira Leal

Mestrado em Tecnologias de Informação aplicadas às
Ciências Biológicas e Médicas

2009

Resumo

A capacidade de processar e relacionar vastas quantidades e vários tipos de dados é uma das vantagens que as tecnologias de informação e comunicação (TIC) trazem à biologia. Esta capacidade torna-se ainda mais importante quando está em causa o estudo da evolução de sistemas intra-celulares complexos, já que este só se torna possível ao contextualizar correctamente informação de diversos tipos (molecular, morfológica e taxonómica, por exemplo). Neste projecto aplicaram-se TIC na construção de recursos que possibilitam o estudo da evolução de duas características de Eucariotas: o sistema de transporte vesicular e centríolos.

No ambiente compartimentalizado que é uma célula eucariota, o sistema de transporte vesicular permite a movimentação de diferentes cargas de um compartimento para outro, incluindo do interior para o exterior da célula e vice-versa. Este sistema está presente, de forma mais ou menos complexa, em todos os eucariotas, pelo que se assume que também esteve presente no último ancestral que estes têm em comum. Desde então adaptou-se aos diferentes estilos de vida e necessidades do eucariotas actuais.

Para o estudo da evolução do sistema de transporte vesicular é necessário conhecer os perfis filogenéticos dos seus componentes, isto é, é necessário saber em que organismos estes componentes estão presentes ou ausentes. As proteínas do tipo rab são reguladores centrais deste sistema. O objectivo deste trabalho é a identificação e classificação desta família de proteínas num vasto número de organismos que cubram, na medida do possível, a diversidade existente em Eucariotas e a disponibilização destes resultados para a comunidade.

Para cumprir este objectivo, foram utilizadas técnicas de aprendizagem automática e de manipulação de sequências para construir uma ferramenta de anotação automática de rabs. Esta ferramenta, apelidada de Rabifier, actua da seguinte forma:

- selecção de sequências candidatas através da sua semelhança com um conjunto discriminante de proteínas rab e não-rab. A semelhança é medida recorrendo á ferramenta de alinhamento local BLAST;

- confirmação das sequências candidatas através da utilização da ferramenta de detecção de motivos lineares MEME/MAST para identificar motivos específicos das rabs;

- utilização de *clustering* para determinar se novas subfamílias devem ser criadas;
- classificação de sequências candidatas numa subfamília através de modelos representativos de conjuntos de sequências utilizando as ferramentas Psi-BLAST e RPS-BLAST;
- marcação automática dos resultados obtidos como sendo de confiança ou não.

Seguiu-se uma verificação manual das sequências marcadas como não sendo de confiança. No final, de um total de 3058867 sequências de 182 organismos, foram identificadas mais de cinco mil sequências em 182 organismos. O conjunto destes dados permitiu uma análise preliminar de características particulares de proteínas rabs e a predição do conjunto de rabs do ancestral comum dos eucariotas.

Os resultados obtidos foram disponibilizados no website TrafficDB (http://www.igc.pt/trafficdb). Este foi desenvolvido utilizando um back-end escrito em Python com base na *web framework* Django.

A presença de citoesqueleto é outra característica que distingue eucariotas de procariotas. O citoesqueleto é composto por filamentos de actina, filamentos intermédios e microtúbulos. Estes últimos podem criar estruturas do tipo centriolar que são responsáveis por processos essenciais ao bom funcionamento da célula. Por exemplo, o centrosoma actua na segregação de cromossomas durante a divisão celular e tem no seu âmago um par de centríolos. Outras estruturas deste tipo incluem flagelos e cílios, que nucleiam protusões da membrana celular e permitem à célula mover-se ou sentir o ambiente envolvente.

Estruturas centriolares aparecem em todos os principais grupos de eucariotas. Assim, como o sistema de transporte vesicular, postula-se que tenham origem num ancestral comum a todos os eucariotas. Mais uma vez, o estudo da evolução de uma estrutura ancestral poderá conduzir a uma melhor compreensão dos sistemas biológicos actuais. Há descrições de variações do esquema de organização destas estruturas, mas esta variabilidade, essencial para um estudo aprofundado da sua evolução, não se encontra catalogado nem centralizado.

O objectivo deste trabalho é a criação de uma interface web, chamada CentrioleDB, para a anotação de imagens de microscopia electrónica de estruturas centriolares. Esta anotação é feita manualmente utilizando um dicionário controlado desenvolvido por especialistas no estudo destas estruturas.

Para além de possibilitar a anotação de imagens, a CentrioleDB também permite a visualização de imagens já anotadas num contexto taxonómico e molecular. Um utilizador pode rapidamente descobrir em que organismos é que uma determinada estrutura aparece, que proteínas foram experimentalmente mapeadas nessa estrutura e quais os seus ortólogos. Esta é a primeira fase de um projecto de colaboração

com grupos ligados à comunidade centriolar. À medida que os dados de anotação ficam mais completos, poderemos comparar perfis de estruturas e moléculas e fazer previsões sobre que moléculas têm funções relacionadas com estas estruturas.

A implementação da CentrioleDB foi feita utilizando uma base de dados relacional e a *web framework* Django. Foi necessário desenvolver:

- uma estrutura de base de dados que aceite facilmente mudanças e acrescentos ao dicionário controlado de anotação, uma vez que este encontra-se em permanente desenvolvimento e actualização por peritos na área.
- um *backend* que lida com o *upload* de ficheiros de imagens e mantém a ligação entre estas e as respectivas anotações.
- uma interface que permita a anotação e a visualização de informação de uma forma intuitiva para o utilizador.

Neste momento a CentrioleDB encontra-se em funcionamento em http://www.igc.pt/centrioledb e disponível a um grupo restrito de utilizadores por razões de copyright.

As duas ferramentas aqui descritas têm em comum o facto de propiciarem às respectivas comunidades um local dedicado ao estudo da evolução dos respectivos sistemas. A integração de informação relevante com taxonomia contextualiaza-a de uma forma que facilita uma visão global e abrangente da evolução destes sistemas essenciais a todos os eucariotas.

PALAVRAS-CHAVE:
Evolução em eucariotas; anotação automática de proteínas; proteínas rab; manipulação de sequências; anotação manual de imagens; centríolos; desenvolvimento de interfaces *web* de bases de dados.

Abstract

The ability to associate and process a vast amount and various types of data is an advantage that information and communication technologies bring to biology. However, most bioinformatics either focus solely on evolution, and we call it phylogenetics, or ignores the evolutionary history of its object of study. In this project these technologies were used to build resources to facilitate the study of the evolution of two Eukaryote defining characteristics, the vesicular trafficking system and centrioles, by integrating familiar or hitherto unexplored types of data (sequences and electron microscopy images, respectively) with taxonomic information so as to give the data a context from which evolutionary studies of complex systems can be achieved.

To study the evolution of the vesicular trafficking system it's necessary to know the phylogenetic profiles of it's components, which is to say the organisms in which the components are present or absent. Rab proteins are central regulators of this system. One of the objectives of this work is the identification and classification of this protein family in a vast number of organisms that cover, as far as possible, the diversity in Eukaryotes and to make these results available to community. To this end, machine learning and sequence manipulation techniques were used, leading to the identification of more than five thousand sequences in 182 species. The identified sequences are available on the TrafficDB website (http://www.igc.pt/trafficdb).

The work developed to facilitate the study of centrioles consisted mainly on the development of an online interface for the annotation and storage of electron microscopy images and of a controlled vocabulary to facilitate this annotation. This interface, CentrioleDB, has the final objective of cataloguing the different morphologies that centriolar structures can have. It was implemented using a relational database and the Django web framework. At this moment it is functioning at http://www.igc.pt/centrioledb and available to a restricted set of users due to copyright reasons.

The two works developed here use different techniques to obtain their data. The one dealing with rab proteins is based on automatic sequence annotation while the one dealing with centrioles is based on tools for the manual annotation on images. Where they cross is in their final purpose, the study of evolution of complex systems, and in the way the data is presented to the public, always with an eye on evolution, using taxonomy as its proxy.

KEYWORDS:
Evolution in eukaryotes; automatic annotation of proteins; rab proteins; sequence manipulation; manual annotation of images; centrioles; development of web interfaces for databases.

Acknowledgements

This work and the time during which it was done wouldn't have been the same without the participation, input and friendship of other people. For that reason, I wish to thank:

my thesis co-advisor, Dr. José Bártholo Pereira-Leal, for his guidance during the entire project and for creating a very positive work environment at the Computational Genomics Laboratory;

Dr. André Osório e Cruz de Azerêdo Falcão, for his role as thesis advisor;

my colleagues at the Computational Genomics Laboratory, who make it a very pleasant place to work, with special thanks to Paulo Almeida and Renato Alves for making sure everything runs smoothly;

Dr. Mónica Bettencourt-Dias and the Cell Cycle Regulation Laboratory for their enthusiasm and knowledge about CentrioleDB and for their company at the institute;

Joana Pinto and Neuza Matias for their patience and for painstakingly pointing out the bugs in the image annotation interface;

the Gulbenkian Institute for Science for providing the conditions necessary for this work and Fundação para a Ciência e Tecnologia (FCT) and Família Tavares Cadete (FTC) for funding;

my colleagues and friends, past and present, at the Faculty of Sciences of the University of Lisbon for their company, help and friendship during the past five years;

Barbara Vreede, for her friendship and copious amounts of stimulating caffeine and conversation, but not for introducing me to SET;

my parents and siblings, for their continuous support and love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Bioinformatics in cell biology

Biology has greatly benefited from informatics, especially in today's age of whole genome sequencing. The trove of data that modern sequencing and high-throughput techniques generate could not be analysed or stored without the input from the bioinformatics community. Furthermore, bioinformatics has provided valuable context through ontologies and integrated databases like UniProt or Ensembl. Apart from molecular biology, computer science has contributed to microscopy through image analysis and to evolutionary biology through computational and statistical methods to determine phylogenies.

However, to study evolutionary cell biology, that is, how complex systems evolved in a cellular context, what is needed is the integration of some or all of the above-mentioned types of data, molecules, images and evolution, specific to the system under study.

In this work, I develop two tools to allow the study of the evolution of two different systems in Eukaryotes: protein trafficking and microtubule-based organelles. Each tool has different requirements and so different informatics techniques were used in their construction.

1.2 Protein trafficking pathways

One of the distinguishing features between Eukaryotes and Prokaryotes is the presence of membrane bound organelles in the former. With them cells are able to compartmentalise, specialise and optimise the functions of the different compartments. However, they had to evolve components to coordinate the transport between them. This transport takes the form of membrane vesicles that bud from the origin membrane and are transported to the target membrane, where they fuse with it, releasing their cargo.

1.2.1 Biological details

The last Eukaryotic common ancestor (LECA) is thought to have already had a complex trafficking system, the main components of which form those of the trafficking system of today's Eukaryotes [14]. These components include, among other, vesicle coat proteins, SNAREs and, playing an important regulatory role, rab proteins [35].

In 1987 the first rab, the ras-like protein SEC4, was identified in yeast and shown to play a role in the late stages of the secretory pathway [33]. In 1989, the ras-like protein 3 was identified in the rat brain [7]. We know it by its more common name, rab3. Since then, the total number of rab proteins has risen to 11 in yeast and to over 60 in humans. This variation is observed just in Opisthokonta; if we include other Eukaryote major taxa, we can reach hundreds of rabs in the same organism, as is the case of the unicellular excavate *Trichomonas vaginalis* [8].

The role played by the rab proteins in vesicular trafficking is that of regulators and coordinators [35]. Their molecular function, however, is analogous to a switch: when activated (bound to GTP) rab proteins interact with and recruit effectors who will in turn perform the functions necessary for vesicle budding, transport and fusion. The switch aspect of rabs is intimately related to their tridimensional structure. The Switch regions of these proteins change conformation depending on whether they are bound to GTP or GDP. In the GTP-bound state, the surfaces exposed have the capability of interacting with the effectors [29]. Another characteristic of rabs is an unstructured and variable C-terminus domain, at the end of which are normally prenylated cisteines. These post-translational modifications anchor the proteins to membranes.

Rabs are themselves regulated by various proteins. Guanoside exchange factors (GEFs) exchange GDP bound to rabs by GTP, thereby allowing them to recruit effectors. GTPase-activating proteins (GAPs), on the other hand, stimulate the hydrolysis of the GTP molecule, thus inactivating the rab [35]. Some rab effectors are able to recruit GEFs to their vicinity, preventing the untimely deactivation of the rab to which they were bound [16]. Rab escort proteins (REP) are responsible for presenting newly-synthethised rabs to the enzimes that add the prenyl groups. After the post-translational modification, GDP-bound rabs are recognized by GDP dissociation inhibitors (GDIs). GDIs have the capability of removing rabs from the membranes. Thus, by forming complexes with rabs in their inactivated forms at the target membranes, GDIs bring them into the cytosol and near the origin membranes, where GDI displacement factors (GDFs), break the complex and reattach the prenyl groups to the membrane [28].

Rab effectors do not bind all rab proteins. Instead each rab subfamily has its set of effectors. This, combined with specific cellular localisation, allows each rab subfamily to regulate specific steps in the trafficking pathways. For example, rab3,

the first rab to be identified in mammals, regulates the secretory pathway in neurons. Other rabs are present in all cells, but still have a specific localisation and function, like rab5, which localises to the plasma membrane and regulates the formation of early endosomes, or rab1, which localises to the endoplasmatic reticulum and regulates the transport from it to the Golgi apparatus. This specificity in function and localisation make the rab proteins good candidates for organelle markers. After all, if a specific organelle is missing in a species, it is likely that the rab that controls the traffic to or from that organelle will be missing as well.

The evolution of protein trafficking components, including rabs, is thought to have happened through duplication of ancestral sequences, leading to paralogues that diverged and specialised in new functions [14].

Given their presence in LECA, their important role in a defining Eukaryotic process and their ability to mark the presence or absence of organelles or transport-related processes, knowledge on the evolution of rabs may shed light into the relationships between the various Eukaryotic groups. However, we first need to identify the phylogenetic profile of different rabs in a broad selection of organisms. Until now, most of the organisms with well-characterized rabs are either Metazoa or Fungi. There are some others in further away branches of the evolutionary tree, but they are few and far between. A systematic identification and classification of rab proteins in as many species as possible would create an invaluable resource to the protein trafficking community.

1.2.2 Objectives and techniques used

In this work, I attempt to systematically identify and classify the rab proteins of near two-hundred species. To this end I developed a workflow, nicknamed the Rabifier, to automatically identify and predict rab proteins when given the protein sequences present in a genome. I also used a combination of automatic and manual verifications to validate the predictions of the Rabifier and built a web-based interface to share my results with the community.

When possible, the Rabifier uses already developed tools. Specifically:

1. BLAST was used to measure sequence similarity [2];

2. BLAST variants Psi-BLAST and RPS-BLAST [3] were used to, respectively, build and search position-specific score matrixes (PSSM) describing a set of sequences;

3. the expectation-maximisation algorithm MEME/MAST was used to detect sequence motifs [4];

4. ClustalW [9] was used to do multiple sequence alignments and Neighbour-Joining phylogenetic trees.

The workflow implementation was done using Python and its BioPython packages [11]. The results were stored in a relational MySQL database. The different steps of the Rabifier workflow were validated by leave-one-out and cross-validation approaches.

TrafficDB, the website that houses the final Rabifier results and which will serve as the basis for a community resource housing information about every component of the protein trafficking pathways, was implemented using the Python-based Django web framework for the back-end and a MySQL relational database for data storage.

1.3 Microtubule-derived organelles

In addition to the membrane-bound organelles, the Eukaryotes are also distinguished from the Prokaryotes by their complex cytoskeleton. It is composed by actin filaments, intermediate filaments and microtubules.

Microtubules are cylindrical arrangements of tubulin. They can serve as rails for vesicle trafficking or organise themselves into bigger structures that have essential and varied functions inside the cell. These microtubule-derived organelles are postulated to have a common origin, but the study of its evolution is dependent of a prior assessment of the various morphologies underlying its functions.

1.3.1 Biological details

Ever since Antonie van Leeuwenhoek and Robert Hooke observed the first cells with a microscope in the seventeenth century, images and the information contained therein has been a part of cell biology. The advent of new microscopy techniques only highlighted the great morphological diversity among living organisms. However, the focus of molecular biology on model organisms, although providing in-depth information about a few tips of the evolutionary tree, did not help us understand how conserved are our findings when moving across species.

Centriolar-like structures are microtubule-based Eukaryotic organelles that act in cell division, chromosome separation, cell motility, cell sensing and transport inside the cell (acting as microtubule organising centers) in their various guises as centrosomes, axonemes and flagella. In some form or other, these structures are found in all crown Eukaryotic groups. While not all branches have them, like higher plants and yeasts, this fact is better explained by secondary loss than by the same structure arising independently several times in evolution.

The overall structure if these organelles is conserved, consisting of a cylindrical arrangement of microtubules, but there is plenty of variation inside this main

organisation. The fold-symmetry can change from species to species, for example. The axoneme of a paramecium has nine symmetric microtubules, while some wasps have hundreds. Some cells have only one cillium, others have many. Structures associated with the cylindrical organelle can vary between species or within different centriolar-structures in the same species.

The molecular biogenesis of this organelle is still under study, but several proteins are know to be involved, including SP2/CEP192, SAK/PLK4 and SAS6 [13]. However, to fully understand the evolution of centrioles, we need to look past individual molecules and into the morphological variation that is the result of the molecular activity. The data to assess this variation exists in the form of decades of exquisite and detailed electron microscopy (EM) images of centriolar-like structures. However, this data is not annotated, centralised or placed in comparison with the recent molecular-based discoveries on centriolar biogenesis.

1.3.2 Objectives and techniques used

In this work, I build CentrioleDB, a community resource to address the lack of integration of decades of electron-microscopy data with today's molecular biology knowledge. CentrioleDB is a web-based EM picture annotation and retrieval interface that also supplies molecular information of interest to centriolar structures, all put in their proper taxonomic (and, by proxy, evolutionary) context.

To serve its purpose, CentrioleDB has the following requirements:

1. an evolving controlled vocabulary to properly describe microtubule-derived organelles;
2. a database schema that allows image annotation using the controlled vocabulary;
3. a web interface that allows users to annotate and upload electron-microscopy images from the literature;
4. said web interface must also allow users to retrieve previously uploaded images and stored molecular information.

The database was implemented with the database management system MySQL, while the website uses the Django web framework as its back-end.

1.4 Thesis structure

The work here presented can be organised along two different lines: the biological backgrounds or the informatics techniques used. Should the latter line of organisation be chosen, the work is separated between a machine learning part, regarding

the automatic annotation of proteins, and a database and interface building part, regarding the manual image annotation and the presenting of results to the cell biology community. If structured along the biological backgrounds, the work is instead separated into a part regarding rab proteins, which includes machine learning for the automatic annotations and database and interface development for presenting these annotations, and a part regarding centrioles, which consists mainly of database and interface development for manual image annotation and retrieval.

I decided to use the biological background line of organisation to present my work, separating it in two chapters:

1. one describing the work done on rab proteins, which was the development of the Rabifier and of TrafficDB;

2. another describing the work done on centrioles, which was the development of the CentrioleDB image annotation interface.

A third chapter in the end provides the general conclusions of the work here presented and its future directions.

I feel that this structure will produce greater clarity when introducing the problems that need to be solved and the results obtained. In addition, the machine learning techniques are used exclusively on the rab protein theme and, while there is some overlap with website and database development on both biological themes, most of that kind of work regards centriole image annotation. The separation on biologial contexts thus also provides, to some extent, separation on the informatics techniques used.

Chapter 2

Rab proteins

2.1 The problem: how to classify?

Each Rab subfamily has a specific location and function in the cell. Sometimes two rabs will regulate different movements between the same organelles. For example, the rab1 subfamily coordinates movements from the endoplasmatic reticulum to the Golgi, while the rab2 subfamily regulates vesicle trafficing in the opposite direction. Taking into account that a rab only acts through its effectors, its function specificity is a result of its effector binding specificity.

It should be stressed here that a subfamily is more than a arbitrary group of rab proteins. Members of a subfamily are linked by sequence similarity and function. If we annotate a protein as a rab, we are stating that it acts as a molecular switch in trafficking pathways. If we further specify that it is a rab5, we are assigning it a very specific location (at the cellular membrane) and function (regulate the movement of early endosomes) and that it interacts with the same effectors as others rab5 proteins. Function follows effector binding and effector binding is driven by the sequence. Or, conversely, we can use the sequence as a proxy for predicting to which subfamily a rab belongs and from the presence of a subfamily infer that a specific function is present in an organism.

Two main problems arise when trying to identify and classify rab proteins based on their sequence. One is that its overall similarity with other Ras superfamily proteins may result in Ras, Rho and Arf proteins being incorrectly classified as Rabs. On the other hand, if the objective is to classify proteins to the subfamily level, differences in the sequence of different subfamilies must be found that will help distinguish them.

2.2 The rab protein sequence and subfamilies

2.2.1 As a whole

As described in the introduction, the rab protein is globular in its N-terminus and unstructured in its C-terminus. The organisation of different Ras superfamily proteins shows that the overall sequence organisation is conserved across families, with a six-stranded beta sheet and five alpha helixes.

A comparison done in 2000 by Pereira-Leal and Seabra [26] revealed linear sequence motifs shared across all Ras superfamily proteins and responsible for their GTPase activity. Of greater importance to rab classification, the same study detected five linear motifs (dubbed RabF1-5) that are conserved across the rab family and aren't discernible in Ras, Rho or Arf proteins and identified their consensus sequences. The RabF motifs allow us to better discriminate between the rabs and the other ras superfamily members.

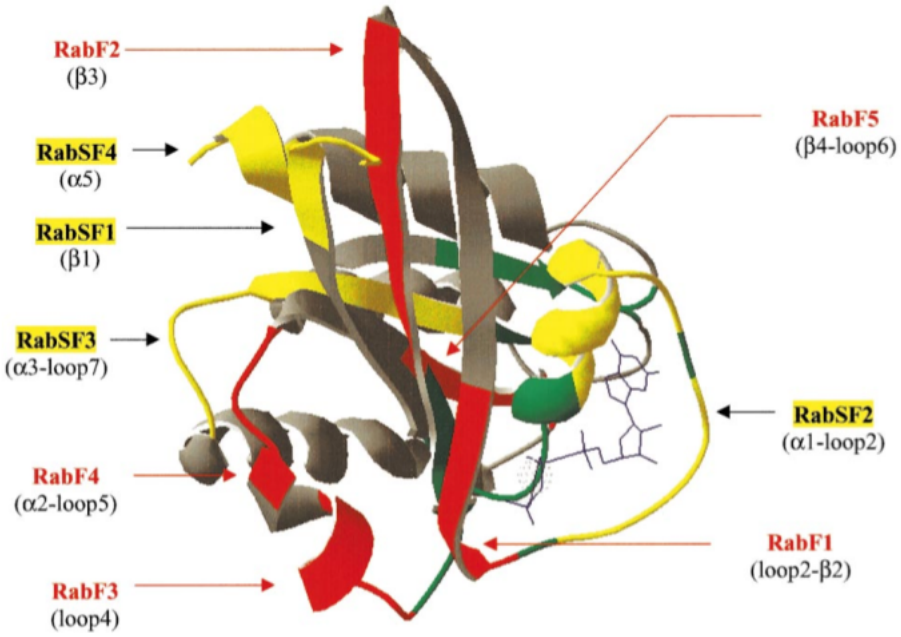


Figure 2.1: Representation of the rab3 tridimensional structure from Pereira-Leal and Seabra, 2000 [26]. RabF motifs are depicted in red, RabSF regions in yellow, and the conserved nucleotide binding motifs in green. The alpha-helixes and beta-sheets were the conserved regions and motifs are inserted are also identified.

2.2.2 Characteristic motifs

In adition to the RabF motifs, Pereira-Leal and Seabra also found four rab subfamily specific regions (RabSF1-4). These regions, while not lying in the Switch regions,

have been shown to act as part of the interaction surfaces, helping establish rab-effector specificity. With their subfamily specificity, the RabSF regions could help distinguish proteins from different subfamilies based on their sequences.

2.2.3 What we don't know

Despite us knowing the constituent parts of a rab sequence, some caveats should apply when attempting to design an automatic classification workflow. One should keep in mind that the previously identified motifs relied on a set of sequences that consisted mainly of Metazoa. While previous studies have show the conservation of RabF motifs in several branches of the tree of life,[27] [1] [20] one should allow for some variation in the consensus sequences.

As regards the RabSF regions, a recent study of two rab subfamilies (rab7 and rab9) across the eukaryotic three of life showed that they are recognisable but not as conserved as one might expect [21]. A further problem in using the RabSF regions in a automated manner is the uncertainty about their precise location and the lack of known consensus sequences for each subfamily.

2.2.4 Goal

Given that identifying a subfamily in a species amounts to stating that a particular step in the trafficking pathway exists in that species, a survey of the rab subfamily profiles on the different organisms that make up the Eukaryotic tree will act as a proxy to describe the trafficking system in those organisms, and, by comparing it across organisms, to describe its evolution. To achieve this end, my goal is twofold:

1. given a genome, identify which of its proteins are rab proteins;
2. assign a rab protein a subfamily and a function based on its sequence.

2.3 The Rabifier

Armed with the knowledge of the rab sequence, I designed a workflow, named Rabifier, to automatically identify and classify to the subfamily level the rabs in complete Eukaryotic genomes.

My objective is to fill in the blanks in rab annotation in the Eukaryotic tree. We currently know a great deal about the rabs from Metazoa and Fungi and we have some knowledge about specific organism scattered around the tree, but for most of the sequenced species there have been no studies. A dataset that covers most of the sequenced Eukaryotic organisms should prove invaluable for future studies on the evolution of this ubiquitous protein family and of the endomembrane system.

To illustrate the scope of this endeavour, most bioinformatics projects on rab classification focus on a single organism [30] [1] [31] [20] [40]. Recently, one project tried to classify rabs in 28 fungi [25] while another covered the full Eukaryotic tree, but only for two rab subfamilies [21]. My project encompasses 182 different genomes and starts with most of the known rab subfamilies, with the possibility of identifying new ones.

Given the vast number of organisms to be classified and the small number of organisms for which we have information, it is not advisable to blindly trust the results of the workflow, even if the validation done yielded good results. Our validation can only be based on the information we have available, which does not span the necessary evolutionary distance. The Rabifier was developed keeping in mind that a good proportion of its predictions would have to be manually validated.

2.3.1 Building a reference set

I do not start classifying the rabs from a blank slate. Previously annotated sequences served as the starting point from which new rabs will be annotated. This makes the Rabifier workflow an instance of supervised learning.

The reference set is comprised of previously [27] and manually compiled sequences from human, *Saccharomyces cerevisiae* and *Caenorhaditis elegans* which are annotated as rab/ypt in Ensembl [18], SGD [10] and Wormbase [6]. In addition, sequences for *Plasmodium falciparum* [30], *Trypanosoma brucei* [1] and *Arabidopsis thaliana* [31] were taken from published organism-specific studies.

This reference set was assembled manually and stored in a MySQL database. The following steps were implemented in Python, using several of the BioPython packages [11], and were applied for each of the 182 genomes in a sequential manner. After a first run, the 182 genomes were run again. This is because, in addition to the manual reference set, the workflow also uses the proteins it annotates as the basis of future annotations. If I had run the genomes through the workflow only once, the first genomes would only have the manual annotations as references, compared to the last one which would have, in addition to the manual annotations, the automatic annotations of 182 genomes. It is also because of this incremental approach that several genomes were not run in parallel. The list of genomes analysed is available in the TrafficDB website.

2.3.2 Finding putative sequences

Description

The sequence database used was Superfamily [38] (as it stood on the 28th of September, 2008). This database was chosen because it includes SCOP [12] protein domain

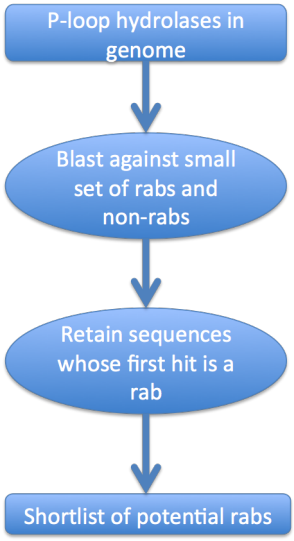


Figure 2.2: Flowchart describing identification of putative rab sequence.

assignments. Being GTPases, all rabs incorporate the SCOP domain 'P-loop containing nucleoside triphosphate hydrolases'. By selecting only the proteins that contain this domain, the number of sequences that must be tested is greatly reduced at no loss of coverage.

For the genome being evaluated I retrieved the sequences that contained the P-loop hydrolase domain domain and ran a BLAST process of each sequence against a discriminating set of sequences containing rabs (yeast and human) and other proteins that, while similar to rabs, belong to other families.

Validation

The discriminating set was refined by repeatedly subjecting *Candida albicans* and *Dictyostelium discoideum* genomes to the SCOP domain selection and subsequent blasting steps. The results were manually analysed by looking at the sequences and their annotations. In each iteration, clearly non-rab sequences that were retrieved were added to the the discriminating set of proteins.

The criterion reached this way to find a putative rab was that the sequence had to find as its best hit a rab with an e-value lower than 10^{-5}. Should the best hit be a rab with a e-value above 10^{-5} or a protein that is not a rab, the sequence being tested is discarded. Different criteria using the top 5 or 10 hits did not yield better results than this simple approach. Through this approach, it is ensured that we select proteins that are more similar to rabs than to any other protein family.

2.3.3 Confirming these putative sequences

Description

After finding a shortlist of putative rabs, the next step was to try and confirm if they were actual rabs. As described previously, rab proteins contain F motifs that distinguish them from other Ras superfamily members. Using the MEME/MAST [4] package I am able to create a probabilistic model for each F motif using the consensus sequence and the rab manual curated reference sequences and use that model to determine if the motifs appear in the putative rab sequences.

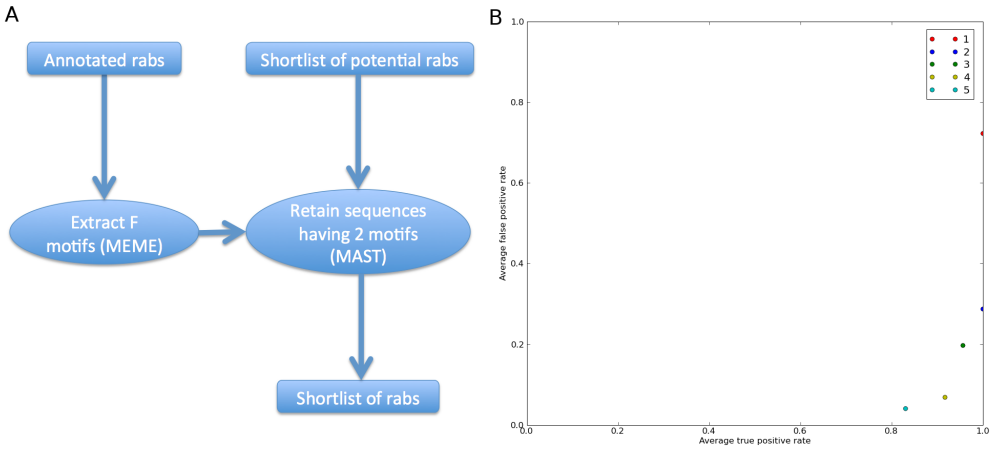


Figure 2.3: A) Flowchart of the F motif-based sequence confirmation; B) ROC curve for the validation of this approach. Each dot represents the average of a cross-validation. The different colours indicate the threshold level used. As can be seen, a perfect true positive rate is achieved with two motifs detected as the threshold.

Validation

This approach was tested using cross-validation by building the model with all the reference species minus one, and then finding the motifs in the species that was left out. In the end, sequences that got through both the first selection and the F motif filter would be compared with the manually curated sequence to determine true and false positive rates. This validation was run for five different threshold levels (for the number of motifs found in the putative sequences). After analysing the results of the cross-validation (Fig. 2.3 B), I chose to use 2 as the number of motifs a putative sequence should have in order to be considered a rab. At this threshold level, the average true positive rate is 1 and the average true negative rate is below 0,3. While it is possible to decrease the latter value at the expense of the former, one must keep in mind that we have a limited reference set. I do not chose a more stringent

threshold to allow for possible variation in some of the motifs when dealing with organisms placed in distant branches compared to our reference species.

2.3.4 Creating new subfamilies

2.3.5 Description

At this point in the workflow we have a list of rabs, but they aren't assigned to any subfamily. Given the number of organisms evaluated and the number of reference organisms, it is likely that some of the rabs to be assigned belong to a subfamily for which we don't have any references. I devised a approach based on clustering techniques to try to automate this process (Fig. 2.4).

We defined the distance $d_{a,b}$ between two sequences a and b as $1 - \frac{N_{sim}}{(L_a+L_b)\times 0.5}$ where N_{sim} is the number of similar residues that align when two sequences are blasted against each other and L_a and L_b are the lengths of each of the sequences. For a sequence being evaluated s and a rab subfamily A containing n sequences $[a_1, a_2, \ldots a_n]$ the distance $D_{s,A}$ between s and A is defined as $\frac{\sum_{i=1}^{n} d_{s,a_i}}{n}$, which is the average distance between the sequence and the members of a subfamily. Each subfamily A has an average internal distance I_A which is the mean of d_{ij}, where i, j make all the possible pairs between the set of sequences in A. Should A have only one sequence, we assume I_A to be the average of I for all subfamilies with more than one sequence.

For each sequence being evaluated s and each existing subfamily A, we check if $D_{s,A} < I_A$. If that is false for all subfamilies, the sequence under evaluation is assigned to a new subfamily. Note that in this step we are no longer comparing our sequence solely to the manually curated reference set, we are also incorporating the sequences previously annotated by Rabifier in the subfamilies when comparing the distances.

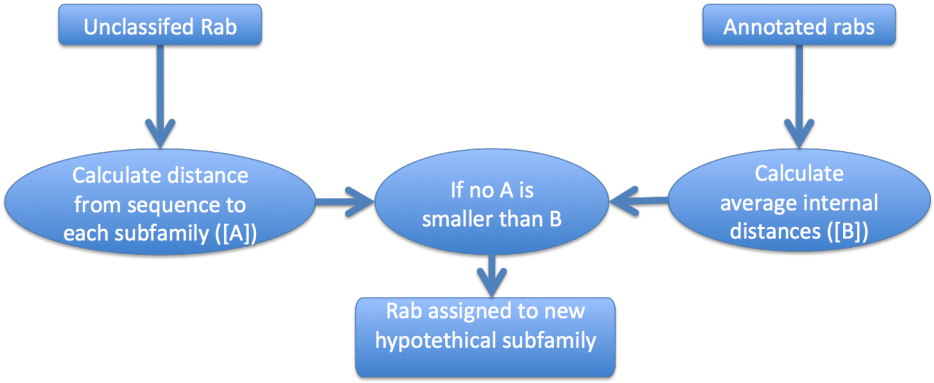


Figure 2.4: Flowchart of the method used to create new subfamilies.

Table 2.1: Validation for new family assignment

	Distance Method		
	Average	Minimum	Maximum
True new families (out of 22)	15 (68%)	8 (36%)	19 (86%)
False new families (out of 175)	49 (28%)	14 (8%)	117 (67%)

Validation

This (admittedly convoluted) method does not guarantee good results. The different taxon-specificity of rab subfamilies means that there are different degrees of divergence between the members of a same subfamily. For example, members of the rab5 subfamily, present in all Eukaryotes, are not as similar to each other as members of the rab3 subfamily, which is present only in organisms with nervous systems. I tested different variations of the described method by taking each of my reference sequences in turn and comparing them with the rest of the dataset. Variation included minimum and maximum distance between a sequence and a subfamily and different thresholds. All gave unsatisfactory results.

Keeping in mind that the workflow results would be reviewed manually (especially the new families), I decided to use the average distance between a sequence and a subfamily and the mean distance inside a subfamily as the threshold, since these parameters balanced the number of true and false new families (Table 2.1).

2.3.6 Assigning sequence to a family

Description

If a sequence does not qualify to start a new subfamily, it must be assigned to an existing family. I used the same technique used by the NCBI Conserved Domain Database [22]. Psi-blast [3] is an iterative version of Blast that returns a collection of related sequences. It also generates a position-specific scoring matrix (PSSM) that describes this collection. A group of PSSMs can be used to create a database that can be queried by sequences with the Reverse Psi-Blast (RPS-Blast) program, much like a Blast database, only it returns a ranked list of PSSMs from the database instead of sequences.

With Psi-blast I created PSSMs for the various subfamilies. As in the previous step, these subfamilies have not only the initial reference set, but also the sequences that the Rabifier has annotated until this point. I then use RPS-Blast to query the PSSMs with the as yet unassigned sequence.

Validation

This was validated using a leave-one-out approach where each sequence in the reference set that didn't belong to a single-sequence subfamily was queried against a PSSM database made with the other sequences. During validation, 160 out of 175 sequences were placed in the correct family.

2.4 Confirming the results of the Rabifier

Due to the number of genomes that we chose to run and the decision not to run several genomes in parallel, the Rabifier took three weeks to pass through every genome twice. The results were stored in a MySQL database between each genome pass to prevent data loss due to unforeseen circumstances.

In total, the genomes searched had 3.058.867 sequences, of which 660.193 were annotated with the SCOP domain 'P-loop containing nucleoside triphosphate hydrolases'. After passing through the discriminating set of rabs and non-rab proteins described in section 2.3.2 and through the RabF motif-based confirmation described in section 2.3.3 the set of P-loop hydrolases dwindled to 6252 sequences identified as rabs.

Most of the sequences identified as rabs were assigned to existing subfamilies, but 653 were assigned to 45 newly created rabHyp (for hypothetical) families. These hypothetical families ranged from unique sequences to a family with 132 sequences.

These results, however, were not yet in a state where they could be trusted. As explained previously, we expected the workflow to produce mistakes, which would be propagated by the inclusion of automatically annotated sequences as new references.

2.4.1 Subfamily alignments

To reduce the number of sequences that would have to be manually analysed, I attempted to automatically flag those that could be wrongly classified. To this end I took advantage of the subfamily-specific SF regions. While I didn't know the consensus sequence or the precise location of these regions in the different subfamilies, they were flanked by the conserved F and GTP-binding motifs. After aligning the sequences of each subfamily (restricting for sequences which were too big and would create long stretches of gaps in the alignment), MEME/MAST was again used to find the F and GTP-binding motifs. Once found, their positions could serve as anchors to automatically extract the rough regions of the SF regions. For each subfamily, the consensus in these regions and its identity to each sequence's SF regions was calculated. Sequences that had a SF region identity with the consensus for the subfamily under a certain threshold were flagged for manual analysis.

To determine the threshold, I once again resorted to my reference set of rab proteins. I randomly added sequences to subfamilies where they didn't belong and tried to flag them using the described method. This was done for identity threshold levels between 0 and 1 at 0.01 intervals, with 1000 random tests per interval. For each threshold, the false positives, defined as sequences that were flagged that were in fact correct, and the false negatives, defined as sequences that were in the wrong family and should have been flagged, but weren't, were counted. The resulting false positive and negative rates are shown in figure 2.5. After analysing these results, I chose 0.65 as the threshold for sequence flagging, as this was a value with a very low false negative rate but still had a manageable false positive rate.

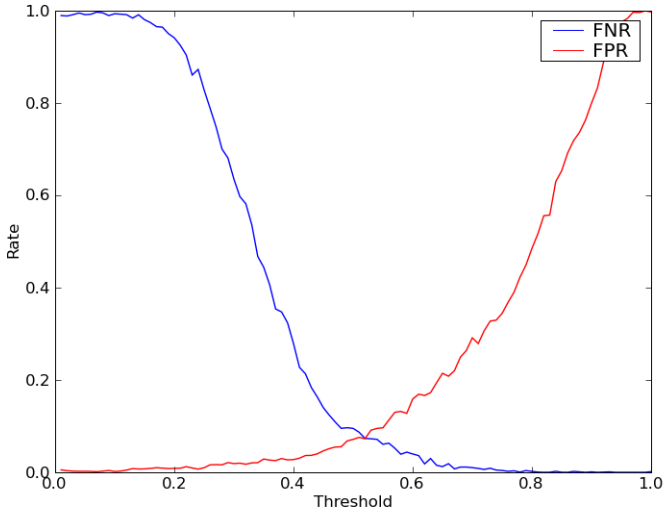


Figure 2.5: False positive and negative rates for correct sequence flagging plotted as a function of the threshold used. Note that 0.65, the value chosen as the definitive threshold, has a very low false negative rate.

After running the scripts, 2471 sequences were flagged. Using the previously calculated subfamily alignments, I manually reviewed which seemed to be in the correct subfamily and which were correctly flagged. Of the flagged sequences, 1019 were removed from their Rabifier assigned subfamilies and marked as unclassified.

2.4.2 Automatic Hyp family assignment

To try to determine automatically which of the rabHyp families were real, two methods were used. The first relied on a 0.7 percent identity between two sequences criterion presented in [27] to determine if two rabs were isoforms. I calculated the percent identity between rabHyps and rabs in other subfamilies. If a rabHyp subfamily had isoforms in a normal subfamily, I would pool the two subfamilies.

For the remaining subfamilies, I tried to detect orthologues between the rabHyp subfamilies and the organisms that constitute my reference set using the bidirectional best hit method [24]. If orthologues were detected, the rabHyp family would be pooled with the corresponding subfamily. In some cases the orthologues were sequences that were not in my reference set as rabs, but were rabs nonetheless, like the human rab45/RASEF and rabL2. This means that the workflow could detect real subfamilies which were not in the original reference set.

2.4.3 Unclassified sequence assignment

The unclassified sequences were manually assigned to subfamilies using a combination of phylogenetic trees for the rabifier results of major taxa and the references, orthology mapping using Blast bidirectional best hits and simple best hits between the unclassified sequences and the reference organisms. The trees were generated automatically by a script which used the ClustalW sequence alignment and Neighbour-Joining program. The Blast queries for the best hits, bidirectional or otherwise, were done using a Python script. On-line databases like Ensembl [18], SGD [10] and Wormbase [6] were used to view the annotations for hits that were not in the original reference set.

A note on nomenclature. The previously established standard is to use letters when defining organism-specific subfamilies, followed by numbers to distinguish among related subfamilies. *Arabidopsis*, for example, has, among others, several rabA subfamilies (named rabA1, rabA2 and so forth). However, the subfamilies defined in different taxa with the same letter may not belong to the same subfamily. In an attempt to use existing names without generating ambiguities, I named letter-defined families after the taxon where they were identified by preceding the letter by the species organism's name initials. This way, the Arabidopsis rabAtA subfamilies can be distinguished from the Trichomonas rabTvA subfamilies. If the taxa where a new subfamily has been discovered is of a higher level than species, more than two letters may be used (as in rabFungiA or rabAlvA, for Alveolata). If a sequence could not be grouped with at least another one to create a subfamily, it is assigned the letter X, as in rabDmX.

Five organisms displayed a great number of unclassified rabs and were analysed more closely: *Trichomonas vaginalis*, *Tetrahymena termophila*, *Paramecium tetraurelia*, *Dictyostelium discoideum* and *Entamoeba hystolitica*. For *T. vaginalis* [20] and *E. hystolitica* [32] previous studies had found and classified their rab proteins, while dictyBase [15] had classification for some of the *D. discoideum* rabs. In these three cases, the sequences recovered by the Rabifier were aligned with human and yeast references and the appropriate organism-specific references. When appropriate, new sequences were assigned to existing families, otherwise they were assigned to new

families. *T. thermophila* and *P. tetraurelia*, on the other hand, were not the object of previous studies. Since both are Alveolata, they were studied together. New families from these two organisms were named rabAlv, rabTt or rabPt depending on whether they contained sequences from both or only one organism.

2.5 Result analysis from rab protein classification

After automatic and manual annotation, 5258 sequences were classified as belonging to one of 147 rab subfamilies, including the non-discriminative rabX subfamily with 42 sequences. The number of sequences in a subfamily varied from two to more than four hundred (Fig. 2.6(a)). Subfamilies ranged from ubiquitous (present in all 182 analysed genomes) to species-specific (Fig. 2.6(b)).

Not all newly discovered subfamilies were exclusive to less studied taxa. Four metazoan rab subfamilies which were not included in the reference dataset were recovered. These are DNAJC27, a protein containg a rab and a DNAJ domain [23], rab45, a protein containing a rab and a EF-hand domain also known as RASEF [34], rabL2 (for rab-Like protein 2) [39] and Partner of ARF (Parf) [36]. As with rab45 and rabL2 in the rabHyp subfamilies, these findings point to the capability of the workflow to discover new bona-fide subfamilies.

2.5.1 Ancestral rab subfamilies

With the rab profiles for organisms that, with the exception of Cercozoa, span all the major Eukaryot taxa as defined by Baldauf in 2003 [5] I can try to see which subfamilies are ancestral. To account for false positive results, I only assume that a subfamily is present in a major taxon if it is present in more than one third of the taxon's species. This is an arbitrary value, chosen to try to avoid situations where a taxa has a very small number of species and, consequently, a small number of wrongly annotated sequences would have great impact. Two groups presented great heterogeneity in the rab profiles of their species and were exploded into more specific taxa. In Opisthokonta, Metazoa shows a great variety of different subfamilies while Fungi has a very reduced set. In a similar manner, the Ciliophora showed a much greater variety of rab subfamilies when compared to the other members of Alveolata, the Apicomplexa.

As seen in figure 2.7, rab 1, 2, 5, 6, 7 and 11 are universal. Rab 4, 8, 18, 21, 23 and 28, while not universal, were present in the ancestral Eukaryote and lost in some of the major groups and this data is supported by both the Rabifier results and the reference dataset. The remaining subfamilies (14, 24, 32, rabL2 and DNAJC27) are shown by the Rabifier results, without support from the reference dataset. If it is confirmed that these subfamilies were indeed present in the last Eukaryotic common

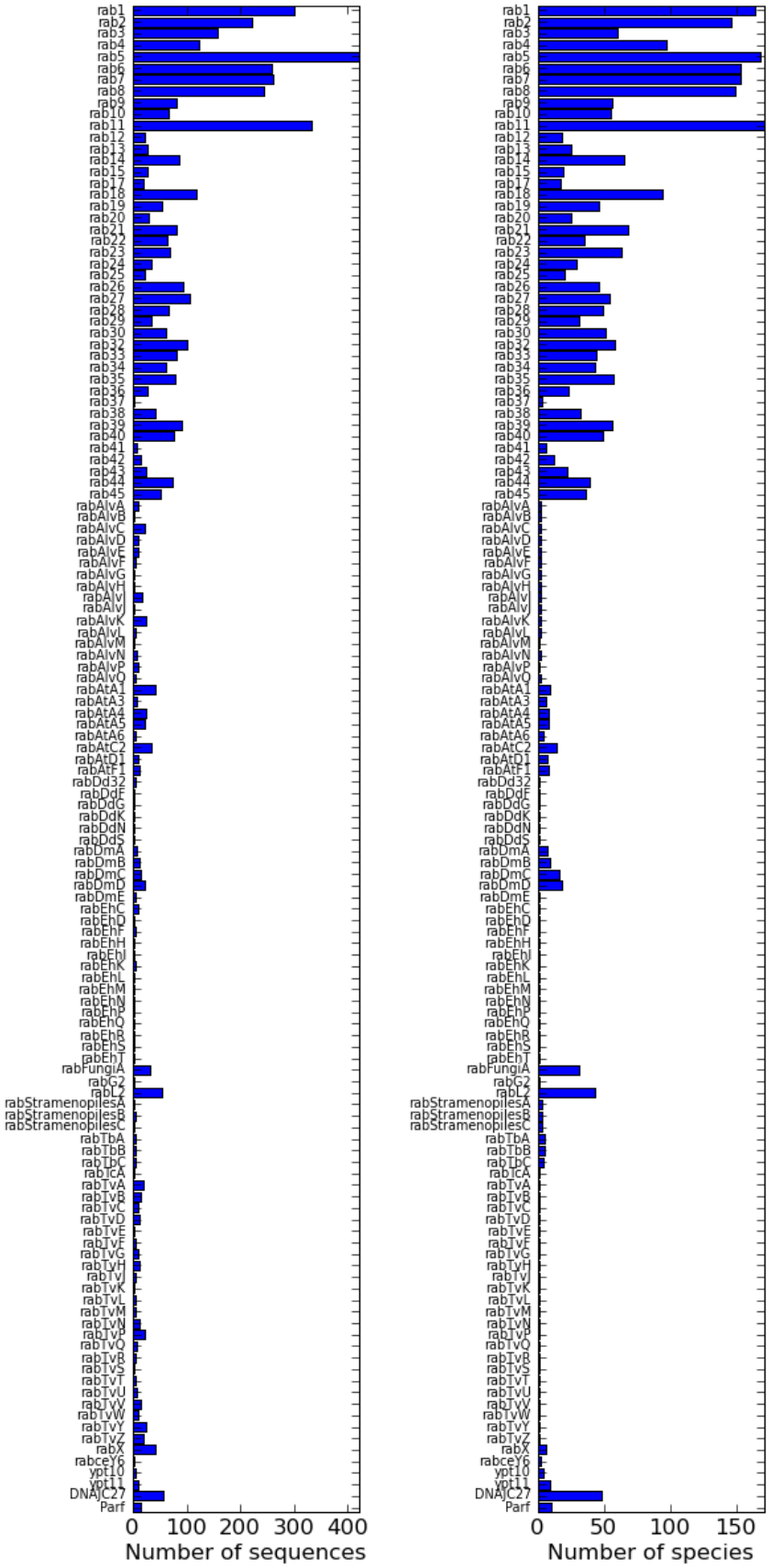


Figure 2.6: a) Number of sequences in each annotated rab subfamily. b) Number of species in which each subfamily appears.

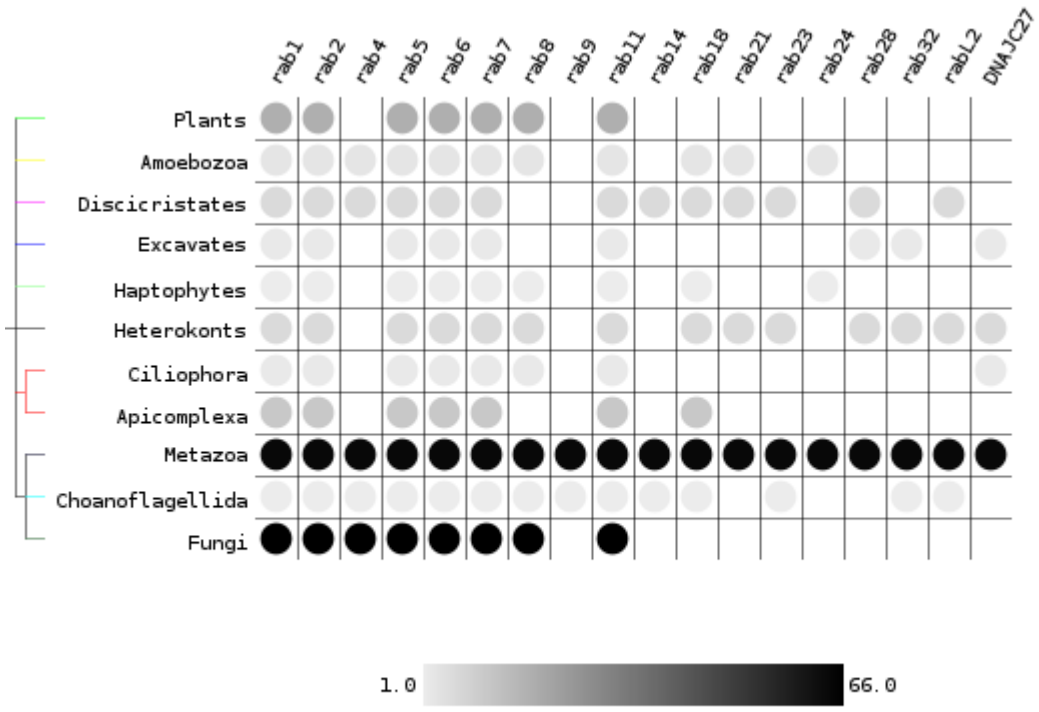


Figure 2.7: Presence of subfamilies in major Eukaryotic groups as obtained by the Rabifier. Only subfamilies that appear in more than one group are shown. The greyscale indicates the number of species available each taxa.

ancestor (LECA), this represents an increase in the variety of its rab repertoire [25] and, by extension, of its trafficking system's complexity.

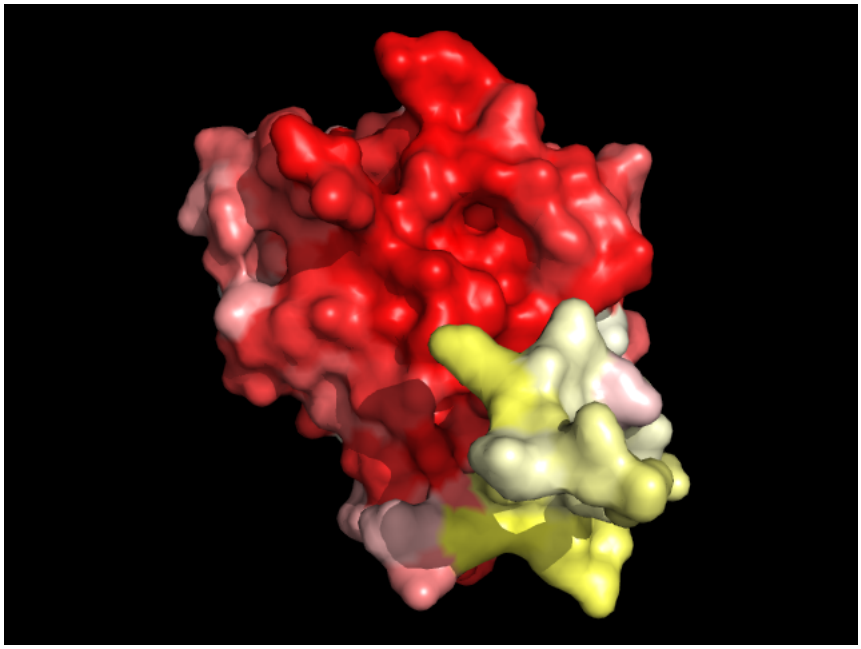


Figure 2.8: Rab1 subfamily sequence conservation mapped on the surface of a human rab1 protein tri-dimensional structure. White denotes lack of conservation, red denotes high conservation. Note the poorly conserved C-terminal hypervariable domain.

2.5.2 Consensus sequences and motif conservation

To evaluate the sequence conservation in a subfamily, I aligned all the sequences belonging to the same subfamily and derived a subfamily consensus sequence. The alignments were also used to map the sequence conservation onto the tri-dimensional structure of the protein. When no structure for a specific subfamily was available, the alignment was mapped to the structure of rab1. In the example provided in figure 2.8 we can clearly see the high degree of conservation (shown in red) of the effector interface and the poor conservation in the C-terminal hypervariable domain. The various conservation-structure mappings are available in TrafficDB, described in section 2.6.

The consensus sequences for the previously described ancestral rab families are aligned in figure 2.9. The nucleotide-binding and RabF motifs are well conserved in and between subfamilies, as expected. As for RabSF regions, while on a general level they were conserved inside a subfamily (with the exception of RabSF4), they are less conserved positions than in RabF motifs. The C-terminus hypervariable domain is very poorly conserved in all subfamilies, including, unexpectedly, the presumable

RabSF4 region. This seems to be in agreement with previous findings for the rab7 and rab9 functional group [21].

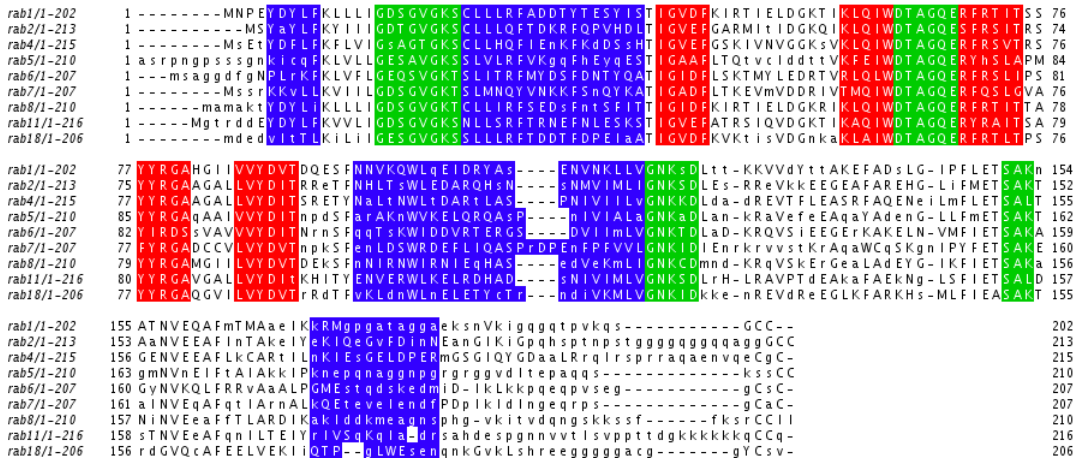


Figure 2.9: Consensus sequence alignemnt. Residues appearing in more than 50% of the sequences are in uppercase. Sequence features are coloured: nucleotide-binding regions are green, RabF motifs are red and RabSF regions are blue.

To further exemplify the conservation of the RabSF regions, figure 2.10 shows an alignment of rab1 sequences from each of the major taxa. Again sequence conservation, albeit not absolute, is observed in the RabSF regions, with the exception of the C-terminal RabSF4.

2.5.3 Rab functional group expansions

Some rab subfamilies are clustered by function and phylogeny into eight functional groups [27]. Using a z-score, I compared the mean number of rabs that each major taxa had from a specific functional group with the average of those means over all the major taxa to determine in which, if any, taxa the functional groups had expanded (Fig. 2.11). Functional groups III (associated with secretory granules) and IV showed a great deviation from the mean in metazoa, and with good reason: they only appear in that taxon. Plants evidenced an increase in functional groups II, VII and VIII, corroborating the findings of Rutherford and Moore in 2002 [31] who presented evidence for the expansion in *Arabidopsis thaliana* of subfamilies associated with rab11, 7 and 8 (corresponding respectively to the functional groups mentioned). Other prominent expansions are those of functional groups I and VII in Amoebozoa and of functional groups I, II, V, VI and VIII in Ciliophora. These taxa include four of the species which had a great number of new subfamilies: *P. tetraurelia* and *T. termophila* in the Ciliophora and *E. histolytica* and *D. discoideum* in Amoebozoa. However, the new families were not taken into account when calcu-

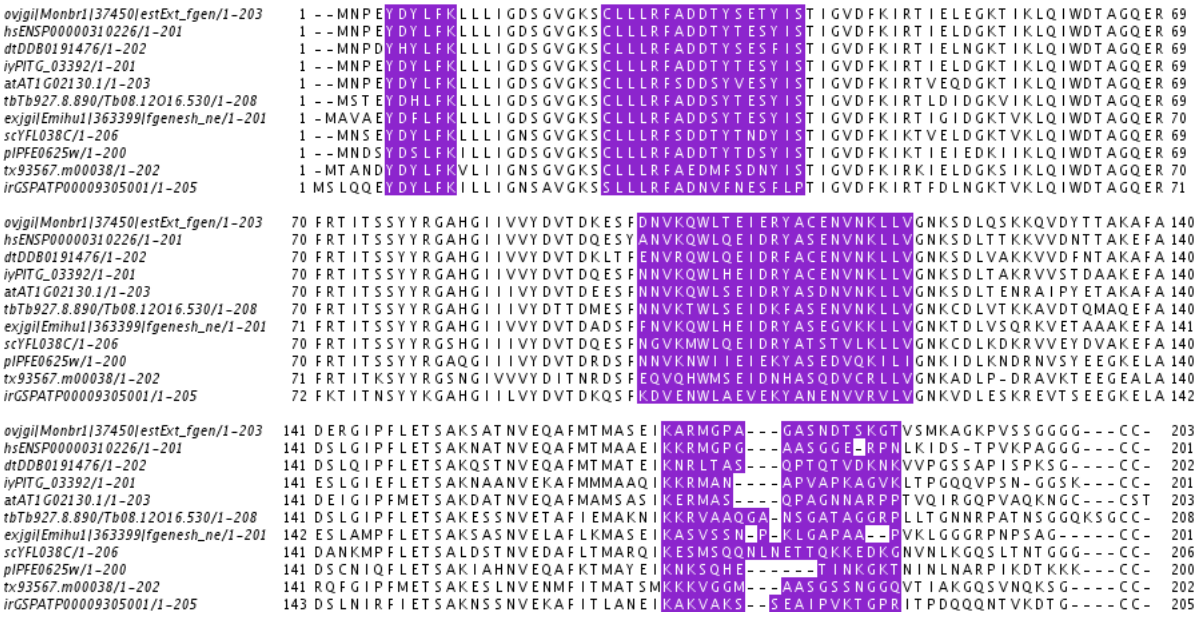


Figure 2.10: Alignment of rab1 sequences from each of the major taxa. The RabSF regions are highlighted.

lating the functional groups expansions; these are due to an increase in these species of the number of sequences of previously identified subfamilies. Three major taxa in particular presented reductions in the number of sequences across the various functional groups: Apicomplexa, Fungi and Choanoflagellida.

A more detailed study of the reasons behind the reduction in the number of sequences in these taxa may shed light on cell biology issues. For example, the reduction in Apicomplexa may be due to the fact that many of its members are intracellular parasites. Choanoflagellida is composed by unicellular organisms and is the closest taxon to Metazoa, to which *Homo sapiens* belongs and which doesn't seem to have suffered any reduction in the numbers of rabs, quite the opposite. By comparing the functional groups that are reduced we can pose questions about the role of the protein trafficking system in multicellular organisms.

2.5.4 Origin of taxon-specific rabs

As mentioned when discussing the assignment of unclassified sequences, five species presented a big number of rabs that did not fit existing subfamilies. Even though these rabs were assigned to new subfamilies, they may have originated from the same ancestral one. If this is the case, it might be that certain ancestral rab subfamilies can adapt more easily to new functions. To evaluate if this is the case, I tried to determine if the taxon-specific rabs evolved from the same subfamily.

By aligning the subfamily consensus sequences and building a Neigbhour-Joining

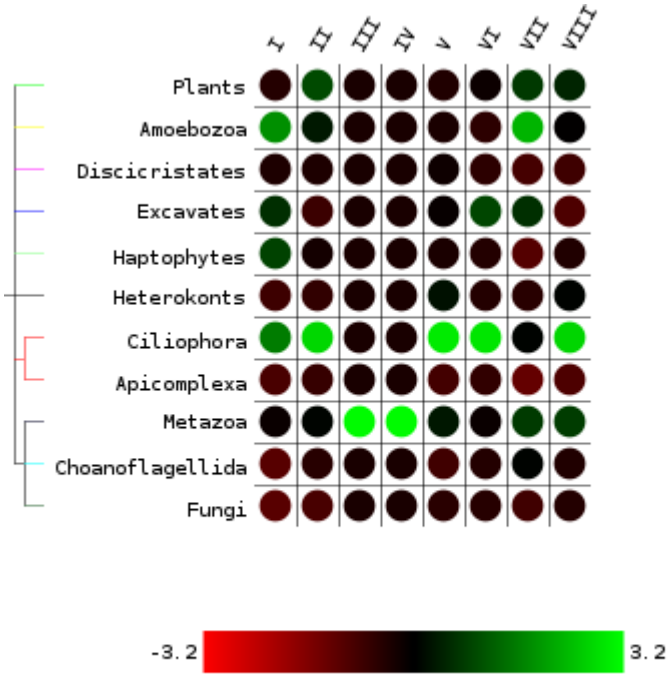


Figure 2.11: Heat-map depicting the number of sequences in each functional group and major taxa as a z-score. Green indicates values above the mean, red values below it.

tree from the alignment, I tried to use the consensus as a proxy for each subfamily and establish the origin of these taxon-specific subfamilies. The resulting tree can be seen in figure 2.12. While, with one exception, the tree of consensus sequences depicts correctly the functional groups, it fails to give any clear indication of which of these groups the new families belong to.

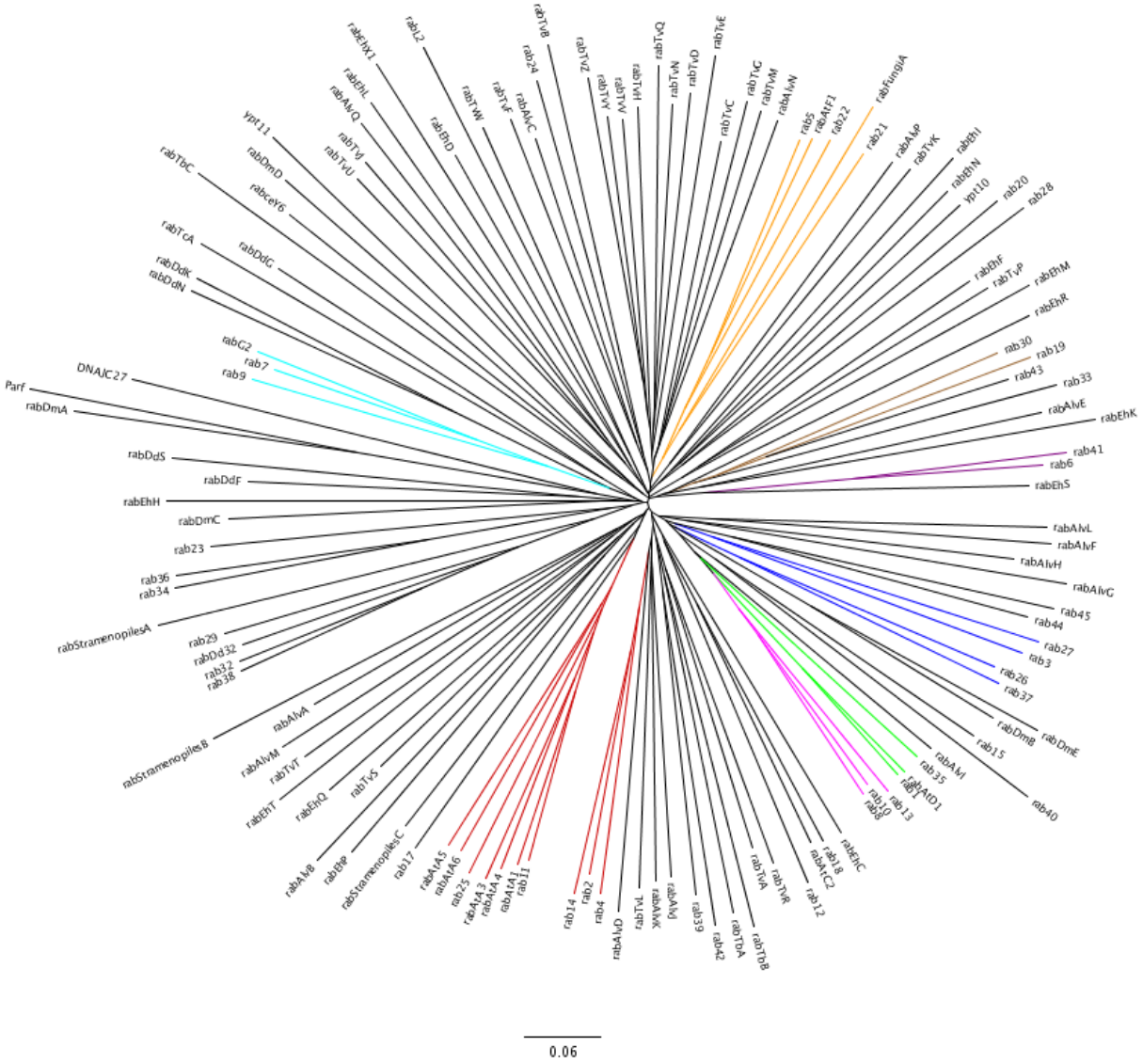


Figure 2.12: Neighbour-Joining tree of the subfamily consensus sequences. The coloured branches represent functional groups. The subfamilies from the five species with a large number of new subfamilies (those starting with rabTv, rabDd, rabEh and rabAlv) did not appear to arise from any of the functional groups.

I then BLASTed the consensus sequences of the new subfamilies against the consensus sequences of the subfamilies which are thought to be present in the LECA. Again, no clear pattern emerged. For none of the organisms did the new subfami-

lies point clearly to one of the ancestors as the more similar. Apparently, in each organism, the duplications that gave rise to the new subfamilies did not arise predominantly from a single ancestral rab.

2.6 TrafficDB

The sequences classified by the Rabifier are of potential interest to the protein trafficking, cell and evolutionary biology communities. Therefore, I designed a website where they are made available. Here users can view them in their proper taxonomical context. In addition, users have access to the subfamily consensus sequences, sequence conservation plots and tri-dimensional rab structures with sequence conservation mapped onto them. The website can be accessed at http://www.igc.pt/trafficdb . While it serves its current purpose of making available the classified rab proteins, the interface is still a work in progress that will evolve as other components of the trafficking pathways are added.

2.6.1 Database schema

The requirements for the database schema of TrafficDB are the ability to store a variety of annotations for a selection of proteins and the ability to connect the proteins to taxonomic information. These two requirements are already fulfilled by the database schema underlying CentrioleDB presented in section 3.2. Therefore, I used the same database architecture on TrafficDB, omitting the picture annotation module of CentrioleDB which isn't needed here.

2.6.2 Website implementation

Like CentrioleDB, the website was implemented using the Python-based web framework Django. Two main views are available for the user, the subfamily and the taxon views.

Taxon view

In the taxon view, users are prompted for a NCBI taxon ID or a taxon name. This does not have to be a species; it can be taxa closer to the root of the Eukaryotic taxonomic tree. Upon selection, users are presented with a list of the rab subfamilies that were detected in that taxon, the identifiers to the sequences in each of the subfamilies and the option to download the sequences in FASTA format.

Subfamily view

The subfamily view provides more detailed information than the taxon view (Fig. 2.13). After selecting which subfamily one wishes to analyse, the user is taken to a page where he can navigate the Eukaryotic taxonomic tree and is provided with simple statistics regarding the number of members of that subfamily present in the adjacent nodes to the one he is in. This way the user can track taxon-specific expansions. In addition to the taxonomic information, the user is also presented with the consensus sequence for the subfamily and with two visualisations of the degree of sequence conservation. One is a plot of the conservation of the most common aminoacid-residue in each position, while the other is a picture of the tridimensional structure of a rab with each residue coloured according to its degree of conservation, as in figure 2.8.

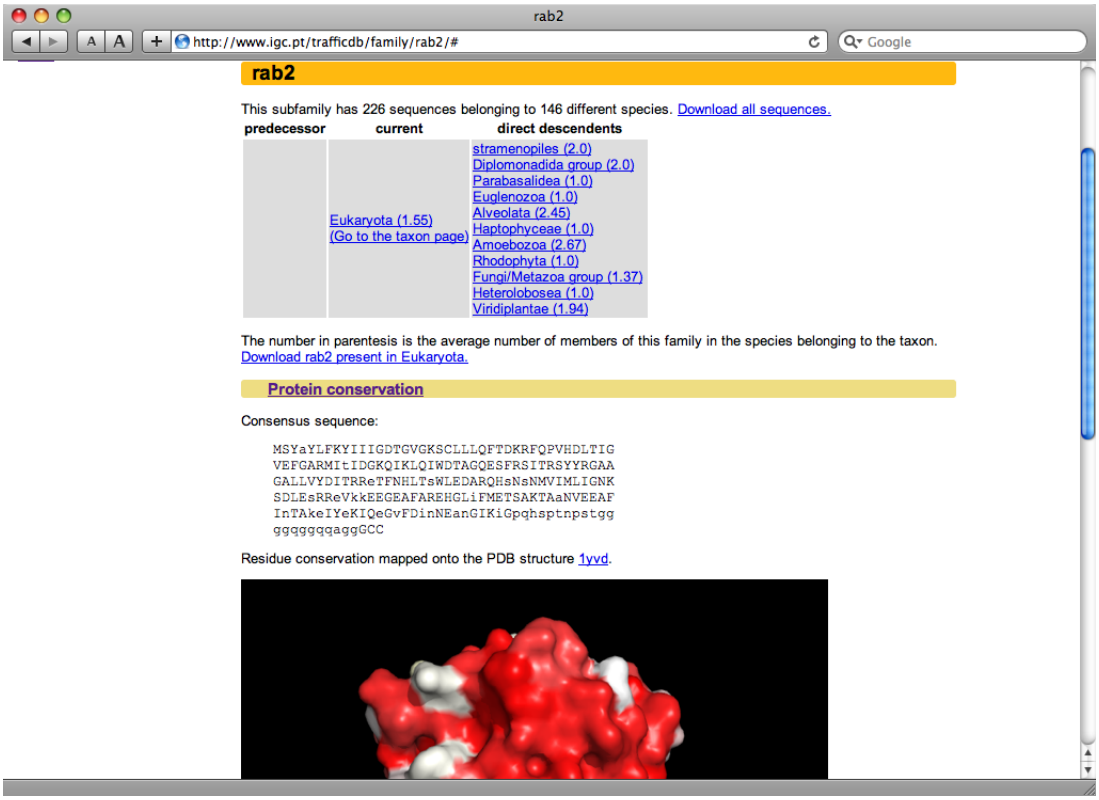


Figure 2.13: Detail of the subfamily view in TrafficDb.

2.7 Future of Rabifier

As mentioned when describing the workings of the Rabifier, this implementation took a long time to run. Therefore, it is not appropriate to use it to classify a single newly discovered sequence or genome. Nor is it needed. The main issue that led to the necessity of this tool was the fact that the previously analysed organisms did

not cover the full taxonomic diversity of Eukaryotes. Now that more species and subfamilies have representative sequences, simpler and more automated methods based on RPS-Blast or Hidden Markov Models can be used, in a similar manner to the NCBI CDD database [22]. In time, TrafficDB will implement this new tool making it a website not only for retrieval of identified rabs but also for identification of new rabs.

Regarding the biological implications of the results presented here, they were subjected only to a superficial analysis. Much more biological knowledge awaits to be mined from it. By studying the expansion of specific subfamilies instead of functional groups or taxon-specific instead of universal subfamilies, one will gain insights into where in evolution did specific cellular processes appear.

Chapter 3

CentrioleDB

3.1 Purpose

The goal of CentrioleDB is to provide the cell biology community interested in centriolar function and biogenesis with a resource that places this structure in its evolutionary context and bridges the gap between morphological and molecular information. Most databases provided to the biology community focus on molecular information: sequences, mutants, markers and so on. By adding morphological information we include an hitherto unexplored type of data in the study of centriolar function and biogenesis.

This resource is not only built for the community but also, in part, by the community, as it provides an image submission and annotation interface using a controlled vocabulary designed to properly describe electronic microscopy (EM) images of centriolar structures taken from the literature. On the molecular sector of the website, users have access to information about the orthology of proteins of interest in centriolar structures, as well as the mapping of proteins to these structures.

CentrioleDB is the basis of a collaborative project encompassing, besides the Computational Genomics Laboratory, experts in the domain of centriolar structures who contributed their knowledge to the development of the controlled vocabulary and have provided and will continue to provide annotations for morphology and sequences. These experts include the Cell Cycle Regulation Laboratory at the Gulbenkian Institute for Science, Professor Keith Gull at the University of Oxford, Michel Bornens at the Institute Curie in Paris and Juliette Azimzadeh at the University of California, San Francisco. Their contribution, feedback and enthusiasm has been invaluable to the project.

3.2 Database schema

The underlying database schema for CentrioleDB is presented in figure 3.1. It needs to fulfil the following requirements:

1. Must support protein and image annotation;
2. New types of annotations must be easily implemented;
3. Proteins and images must have a way to be linked to taxonomic information.

3.2.1 Image and protein annotation

The image annotation module of the CentrioleDB schema has its focus on four tables: 'pics_info', 'picture_annotation', 'picture_has_ann' and 'papers'. 'pics_info' stores the basic information of each image: it's id, file location, figure number, the publication from which it comes (via a foreign key to the 'papers' table) and the taxon to which it belongs (via a foreign key to the 'view_taxon' table). The annotations themselves are stored in the 'picture_annotation' table, with an id, the group to which they belong, the annotation itself and an optional description. For example, the annotation used to describe the image annotation appears in the table as (id = 3; picture_annotation_group = 'image'; picture_annotation = 'image magnification'; description = 'The magnification at which the image was taken.'). The 'picture_has_ann' table implements a many-to-many association between the 'pics_info' and 'picture_annotation' table, with a 'picture_annotation_value' text optional attribute. In the image magnification example described above, this attribute would take the value of the magnification; when associating an image with the annotation that describes the figure legend, the value of the attribute would be the text of the legend.

While this structure does not completely eliminate redundancy, it allows for great flexibility when adding new annotations or new values to existing annotations. The controlled vocabulary used for image annotation changed often during the first months after CentrioleDB was implemented, but the database structure did not have to be modified. Instead, annotations were simply inserted, removed or altered in the 'picture_annotation' table. On the other hand, since the picture annotation value attribute is of the type TEXT, it relies on the previous validation of user input to assure that it is meaningful in the context of its annotation.

The protein annotation module follows the same structure as the image one, only without a 'papers' table and with specific gene/protein tables. It also has a greater number of attributes in the 'has_ann' table, including a foreign key to a table with the different methods used to generate the protein annotations (thus making it a

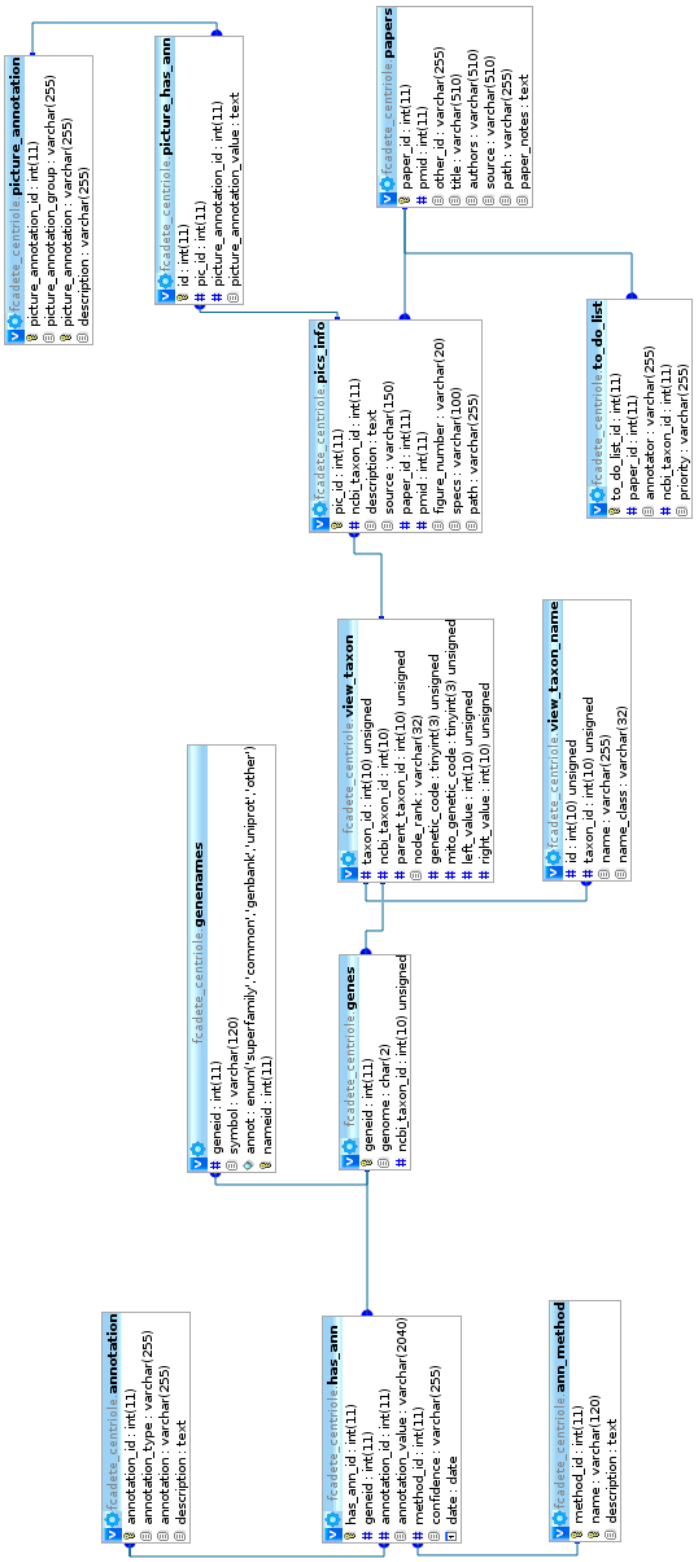


Figure 3.1: Database schema for CentrioleDB. The Django generated tables for user authentication are not shown.

table implementing a triple association, instead of a binary one) and a confidence attribute which takes values appropriate (if any) to the different methods used. The table containing the gene/protein entities (‘genes‘) is also linked to a ‘gene_names‘ table, so that different external identifiers can be used for each protein, like Uniprot and GenBank accession numbers.

3.2.2 Taxonomic information

Both images and genes are connected to a unique NCBI taxon ID. This ID, stored in the ‘view_taxon‘ table, supplies a taxonomical and evolutionary context. The ‘view_taxon‘ and ‘view_taxon_name‘ tables were sourced from the open source project BioSQL and implement the NCBI taxonomic tree. Not only do they store the level and names of each taxa and what is its predecessor in the tree, it also includes two columns with the pre-computed left and right values of a pre-order depth-first transversal of the tree, starting at its root. This way, complex queries involving tree transversals are simplified.

3.3 Development of the controlled vocabulary

The controlled vocabulary used to describe EM images of centriolar structures consists of a collection of specific terms to define different characteristics observed in the images. Some of these terms denote the presence of a main centriolar structure (’centriole’ or ’basal body’ for example) while others refers to details of certain structures (’basal body cartwheel’ or ’axoneme radial spokes’, for example).

Each term in the vocabulary is assigned to an annotation group. To use some of the previous examples, terms like ’centriole’ or ’basal body’ belong to the annotation group ’structure’ while ’basal body cartwheel’ and ’axoneme radial spokes’ belong to the annotation groups ’basal body’ and ’axoneme’, respectively. The groups help define a form of hierarchical association between the different terms, as a image will only be annotated with structure-specific terms if it has been annotated with the term that denotes that structure to begin with. A selection of some of the terms in the vocabulary and their relationships can be seen in figure 3.2.

The vocabulary was developed in an iterative manner. The image annotators started with a set of terms that described the most common centriolar structures. As their worked progressed meetings between the annotators, the domain experts and the database developers were held. In these meetings the annotators would present images where there was doubt about whether some structures could be described using the existing terms and the parties involved would discuss if new terms were necessary. Should that be the final decision, the terms would be added to the database, a process made easy by the database schema used.

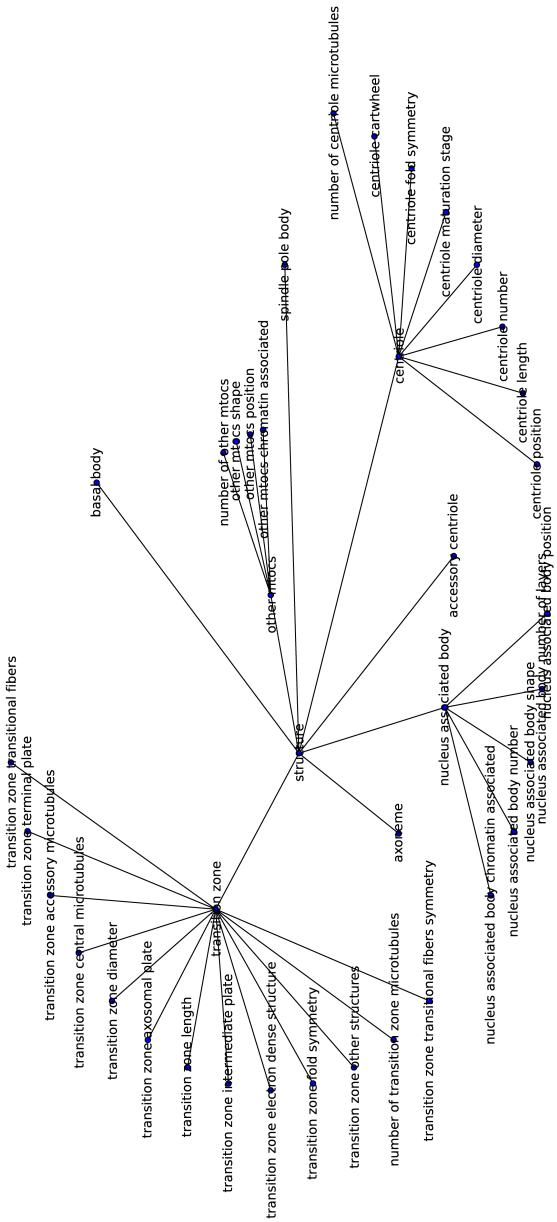


Figure 3.2: Representation of part of the centriole controlled vocabulary. Each node represents a term and an edge means that a term belongs to a group. For example, 'transition zone' belongs to the 'structure' group and the terms characterising a transition zone belong to the 'transition zone' group.

3.4 Website implementation

The website for CentrioleDB was implemented using the Python-based web framework Django. This framework uses specially designed classes representing the database schema, allowing for complex querying inside the Python code. Django also has modules to deal with user authentication, sessions, forms and pagination, for example, facilitating the website development. While the back-end for the system is in Python, Django uses HTML templates with special markup to build the responses to the user.

A note on style. During the development of CentrioleDB I built the HTML templates and corresponding CSS and it was in that form that it first launched. Recently, Marg Gouw, a colleague at the Computational Genomics Laboratory has started to participate on the project and made great contributions to the CSS and to the aesthetic component of the website, so that the way it looks today is the result of our collaboration. On the other hand, the choice of views available to the user, the content that is displayed in each view, and the back-end that generates it, were done by me.

The main users of the website have been Dr. Mónica Bettencourt-Dias, Zita Carvalho-Santos, Joana Pinto and Neuza Matias. They contributed generously with interface suggestions leading to better usability.

3.4.1 From user to database

The Django framework provides easily customised Python scripts and classes to deal with all the necessary steps between a HTTP request by a end-user and the HTML response to be rendered by his browser (Fig. 3.3).

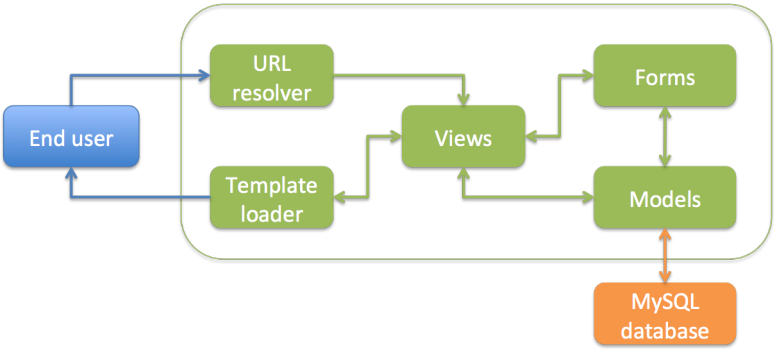


Figure 3.3: Diagram of the inner working of a Django-based back-end. The end-user is represented in blue, Django classes or functions are represented in green and the MySQL database is represented in orange.

When the server receives a request, the first thing Django does is resolve the

URL and call the appropriate function. In Django, a function associated with a URL is called a 'view'. Apart from GET or POST arguments, the views functions can also take arguments encoded in the URL and retrieved using regular expressions in the URL resolver.

It's in the view functions that most of the request and data processing takes place. It does not interact directly with the database, relying instead on 'model' classes. These are defined by the developer and there is usually a one to one correspondence between a model class and a table in the relational database. Restraints on the SQL code and foreign keys can be defined in the model class. This way, when defining the view functions, the developer has Python objects that represent his database and allow him to do any necessary query, insertion or update without having to explicitly mix Python and MySQL in the same functions.

After retrieving and processing the necessary information, the view function supplies context to a template. The template consists mainly of HTML code with specific Django template language that allow the template loader to dynamically prepare the final HTML page. The Django template language implements intentionally simple loops and logic checks, allowing for limited final data processing. After rendering, the final HTML page is sent to the user as the response to his original request.

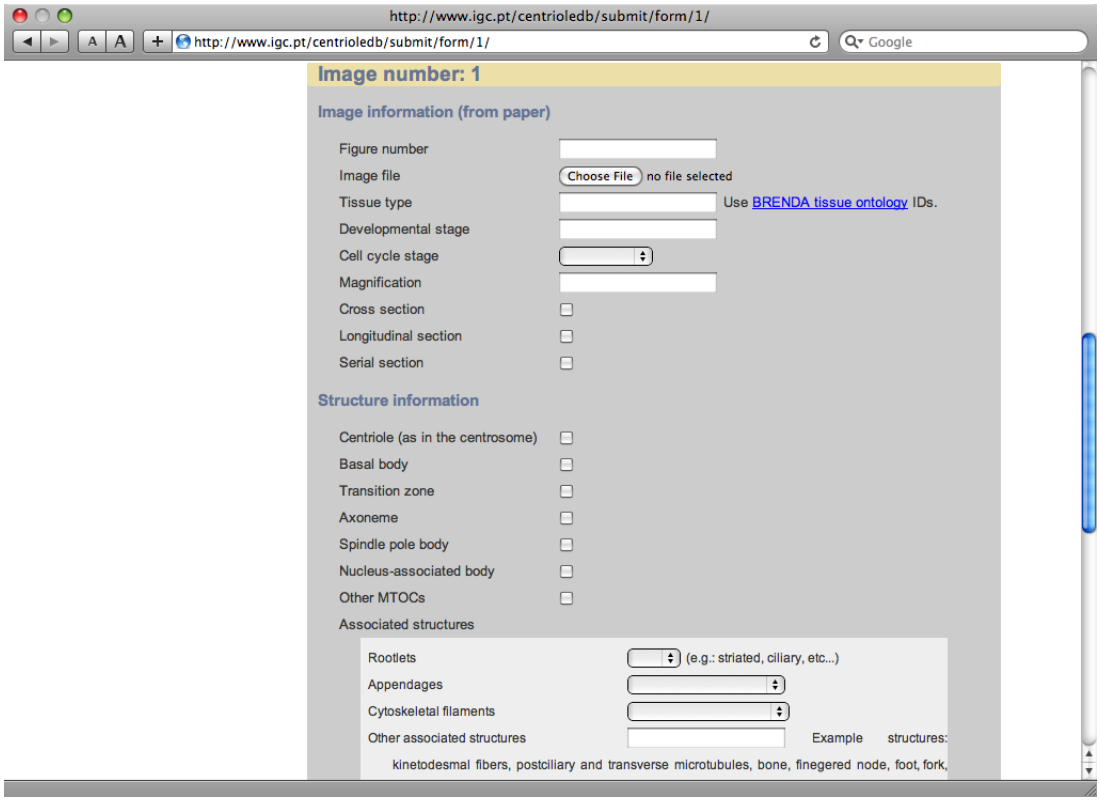
3.4.2 Interaction scenarios

Submitting or editing an image

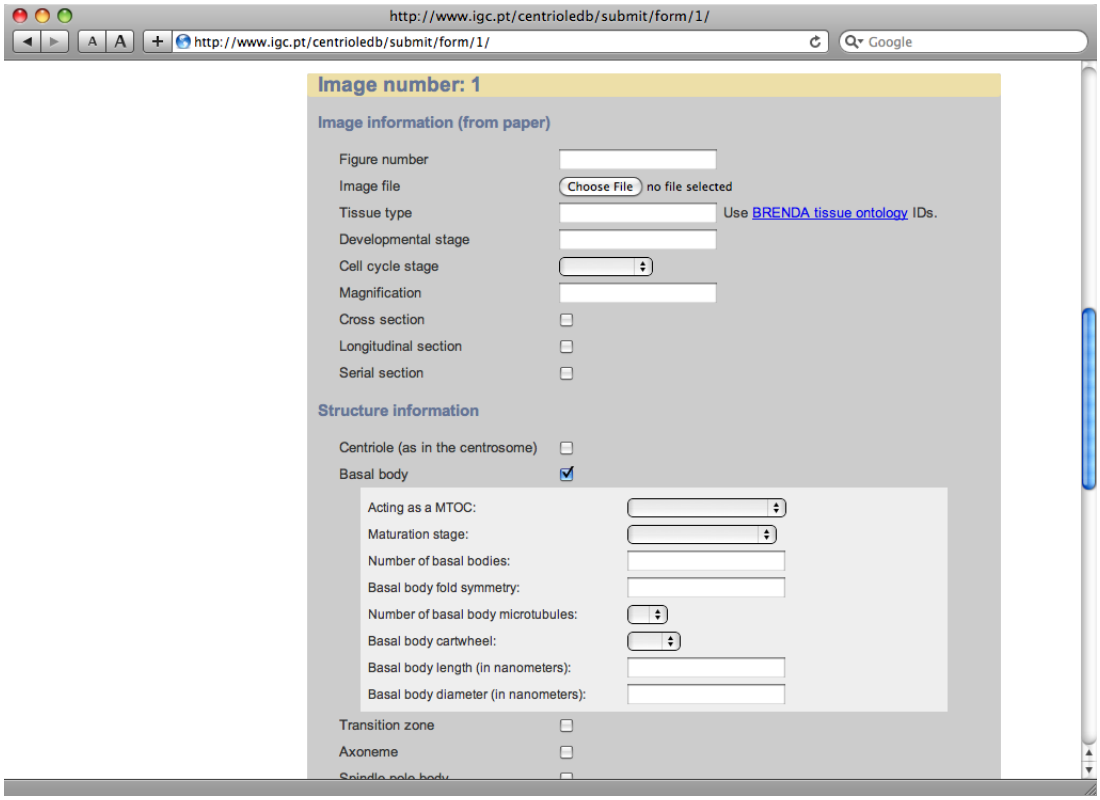
To submit an image the user first clicks on the corresponding link on the left-hand side menu. He is prompted for the number of pictures he wants to submit in one go and, upon answering, is directed to the main submission form. Here the user fills in the necessary information relating to the paper and to the image. Regarding the paper, should it have a PubMed ID, it and a pdf file are all that is required for the complete paper information to be added to the database. Otherwise the user has to manually insert the paper's details.

Regarding the image annotations, the user is presented with a series of check boxes, drop-down menus and text fields with the appropriate labels. In an attempt not to overwhelm the user with choices, some options only appear after certain selections. For example, detailed annotations describing a basal body only appear after the user has indicated that such a structure is present in the image (Fig. 3.4).

Behind the scenes, the form is generated using costumized Django 'Form' classes and templates. In the classes, I defined which fields the form will have, their types, the permissible data and, where appropriate, the available choices. I mentioned when describing the database schema that, to allow for annotation flexibility, the onus is on the user to make sure that he is inserting the correct annotation values, if any.



(a)



(b)

Figure 3.4: Detail of the image submission form. Note the expanding basal body annotations in (b).

In CentrioleDB, the database user is the Django back-end, not the end-user, so by correctly designing the form classes the database and website administrator ensures that the correct type of data is inserted. On end-user submission, the back-end validates the form against the defined classes and should something not be correct (the image file is in fact a text file, for example) returns the filled-in form with the corresponding error messages. On submission of a form that passes validation, the paper and image files and information are inserted into the database and are available for querying right away.

Should a user want to edit the annotations for an existing image, by following the link on the left-hand side of the page he will be presented with the list of submitted papers and, after choosing one, with the list of annotated pictures in that paper, where he can chose to delete one of them (not without being prompted if he is sure about his actions) or simply edit the annotations. In case the user wants to do the latter, he will be directed to a form similar to the original image submission one, already filled in with the existing annotations.

Searching for images

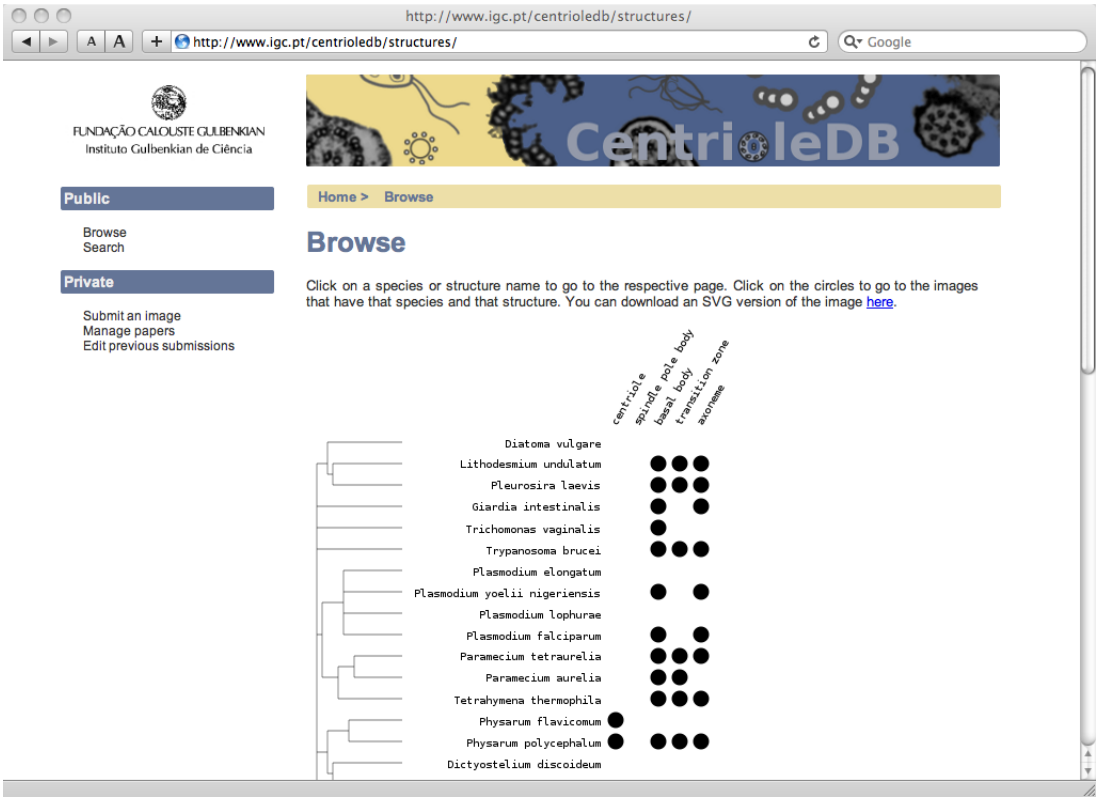


Figure 3.5: The Browse view. The presence matrix is interactive and dynamically generated.

There are two main ways a user can search for images in the database. The most intuitive one is 'Browse' view (Fig. 3.5). Here the user is presented with a dynamically generated presence matrix listing the species for which images are available and the annotated structures for each of them. The presence matrix is interactive; by clicking on a position in the matrix the user is directed to a list of the images that satisfy his request. A second way to search for images is in the 'Search' view (Fig. 3.8). The user selects the characteristic he wants to search for (at the moment the choices are Paper, Taxon and Structure) and the form seamlessly updates to show the available instances of that characteristic. After choosing and submiting the form, the user is taken to a list of images identical to the one he would get had he done the search through the 'Browse' view.

Figure 3.6: An example of the results in a search for images. Note the URL describing the user location.

The image search result view includes thumbnails for the images, the paper from which the image was taken, the figure number in said paper and a list of the main structures present in that figure, as can be seen in figure 3.6. Note the URL on the figure. It describes the status of the user at that moment. In this example, the user was searching for pictures of basal bodies in *Lithodesmium undulatum* and the URL reflects that by stating that the user is doing an image search for the taxon 59812

(the corresponding NCBI taxon ID) and for the basal body structure. This address is static, it will always direct the user for the list of images stored in the database at that moment that answer this query. Similar URLs are used if the search is based only on paper, taxon or structure. On clicking on the desired thumbnail, the user is finally directed to a view describing the image in greater detail using the controlled vocabulary (Fig. 3.7).

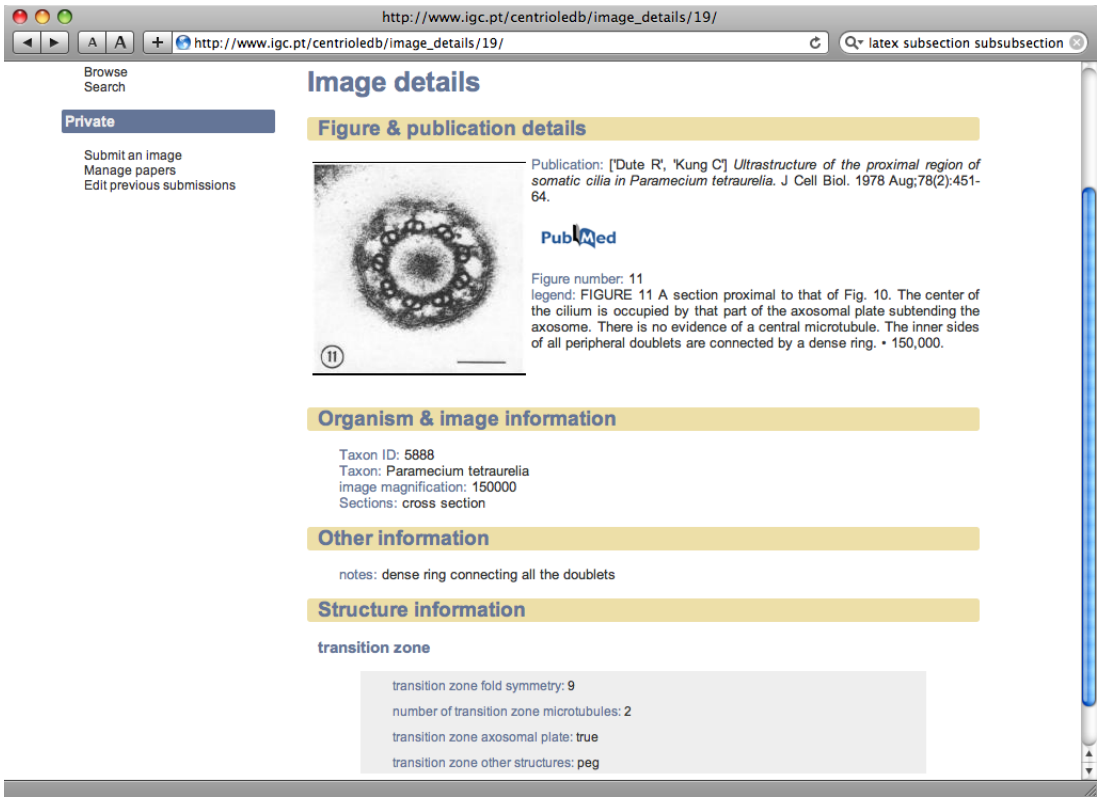


Figure 3.7: The detailed description of a particular image. The vocabulary is the same that is used when submitting the images.

Searching for proteins

The starting point when searching for proteins is the same 'Search' view that can be used to search for images accessible through the left-hand column. However, the user now fills in the lower forms to obtain either a list of proteins that belong to a certain family or proteins that map to certain structure.

Should the user want to search for a family, he will be directed to a view that provides a brief description of the family, the list of sequences that are annotated as belonging to that family and a taxonomic tree showing where this family appears. Statistics about the number of sequences, species and number of sequences per species are also provided. The user can drill-down the taxonomic tree to investigate

Figure 3.8: The general search view, where the user can search for images or for proteins.

whether the distribution of sequences inside a taxon is homogenous.

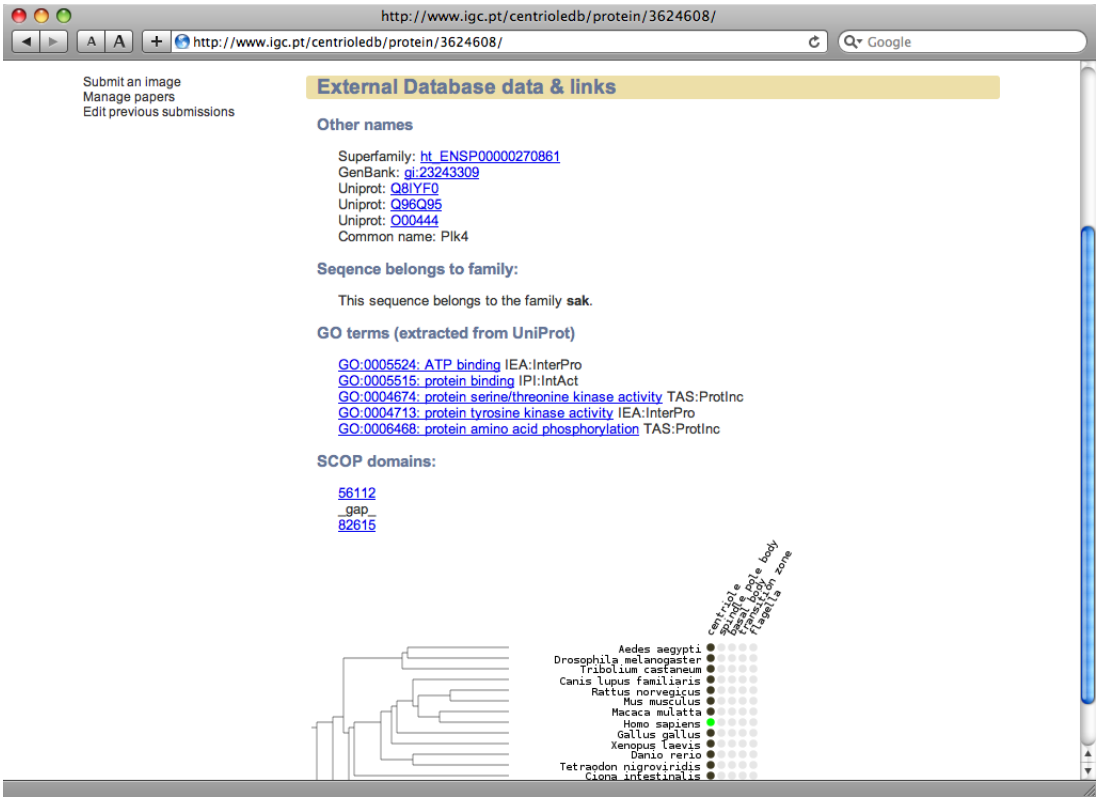


Figure 3.9: An example of the information stored in the database about a particular protein.

If the user searches for proteins that map to a certain structure and selects one of them, or if he selected a protein from the family description, he is taken to a page where more detailed information about the protein is presented to him (Fig. 3.9). This information includes the protein names, GO annotations, families it belongs to, structural domains and a dynamically generated matrix that shows to which structures (if any) the protein is mapped to and what orthologs it has.

3.5 Current status

As of September 6, 2009, CentrioleDB has had 304 annotated images submitted, covering 58 different organisms. The current controled vocabulary has 123 different possible annotations, arranged in 13 different annotation groups, which range from technical image details like magnification to specific characteristics of certain structures, like the presence of a cartwheel in a centriole.

Information about protein mapping to structures came from literature [17] [19] [37] and from a personal communication by Dr. Julliete Azimzadeh. The proteins

associated with families of importance in centriolar function and biogenesis came from work by Zita Carvalho-Santos of the Cell Cycle Regulation Laboratory which has been submitted for publication. Orthologs were mapped for a list of organisms of interest in the study of centriolar structures using the bidirecional best-hit method [24].

As for the future, the ongoing annotation of images will continue, done both by members of the Cell Cycle Regulation Lab and by its international collaborators. In addition, the controlled vocabulary is to be expanded to properly annotate tissue-specificity in multi-cellular organisms and to describe morphological variations arising from mutations. Finally, on the protein side, algorithms are expected to be developed that can deal with the coiled-coil structures that are frequent in centriolar components.

Chapter 4

Conclusions

I expect to proceed with the biological analysis of the protein trafficking network results and to submit a manuscript for publication in October, 2009. In addition, my work during the stay at the Computational Genomics Laboratory also enabled two studies in which I am a co-author and which will be submitted soon:

1. 'Stepwise Evolution of Centriole Assembly Mechanisms', by Zita Carvalho-Santos, Pedro Machado, Pedro Branco, Filipe Tavares-Cadete, Ana Rodrigues-Martins, José B. Pereira-Leal and Mónica Bettencourt-Dias

2. 'Extensive innovation in the evolution of Rab:effector interactions by Maria Luisa Rodrigues, Filipe Tavares-Cadete and José B. Pereira-Leal.

In this work I present two new resources for the cell biology community. Their differences are more than the cellular processes each one covers or the type of information displayed. A fundamental distinction is the way that information was obtained. In the first resource presented here, TrafficDB, the information contained therein was compiled by my own bioinformatics analysis. In contrast, the main component of CentrioleDB, its images, were not a direct result of my work, which consisted of providing the means for users familiar with centriolar structures to upload existing images and annotate them themselves.

There is still scope for improvement in both projects. In TrafficDB, apart from the already mentioned development of a lighter and faster tool to identify rab proteins in new sequences using the results of the Rabifier and further analysis of its biological significance, the next logical step is the incorporation of data on the other components of the protein trafficking system. These include rab effectors and regulators, SNARE proteins and vesicle coat proteins, to name a few. By adding the different components of the system, we can further infer its state in the LECA and how it adapted and innovated, evolving into the forms present nowadays. The TrafficDB system is already designed with this in mind and rabGAPs, while not being displayed in the interface, have already been added to the database.

CentrioleDB will benefit from ongoing annotation of more structures in a greater variety of organisms. When we have sampled the taxonomic and morphologic diversity, we can attempt to correlate it with the molecular diversity, through the comparison of phylogenetic profiles of structures and molecules. If a particular protein only appears in organisms with a particular structure, we can use this as a prediction that the protein is involved in the formation or interacts with the structure. The best of these predictions can then be brought to the laboratory to be experimentally validated.

The two works developed here use different techniques to obtain their data. The one dealing with rab proteins is based on automatic sequence annotation while the one dealing with centrioles is based on tools for the manual annotation on images. Where they cross is in their final purpose, the study of evolution of complex systems, and in the way the data is present, always with an eye on evolution, using taxonomy as its proxy.

Bibliography

[1] John P Ackers, Vivek Dhir, and Mark C Field. A bioinformatic analysis of the rab genes of trypanosoma brucei. *Mol Biochem Parasitol*, 141(1):89–97, May 2005.

[2] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, Oct 1990.

[3] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, Sep 1997.

[4] Timothy L Bailey, Nadya Williams, Chris Misleh, and Wilfred W Li. Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic Acids Res*, 34(Web Server issue):W369–73, Jul 2006.

[5] S L Baldauf. The deep roots of eukaryotes. *Science*, 300(5626):1703–6, Jun 2003.

[6] Tamberlyn Bieri, Darin Blasiar, Philip Ozersky, Igor Antoshechkin, Carol Bastiani, Payan Canaran, Juancarlos Chan, Nansheng Chen, Wen J Chen, Paul Davis, Tristan J Fiedler, Lisa Girard, Michael Han, Todd W Harris, Ranjana Kishore, Raymond Lee, Sheldon McKay, Hans-Michael Müller, Cecilia Nakamura, Andrei Petcherski, Arun Rangarajan, Anthony Rogers, Gary Schindelman, Erich M Schwarz, Will Spooner, Mary Ann Tuli, Kimberly Van Auken, Daniel Wang, Xiaodong Wang, Gary Williams, Richard Durbin, Lincoln D Stein, Paul W Sternberg, and John Spieth. Wormbase: new content and better access. *Nucleic Acids Res*, 35(Database issue):D506–10, Jan 2007.

[7] E Burstein and I G Macara. The ras-like protein p25rab3a is partially cytosolic and is expressed only in neural tissue. *Mol Cell Biol*, 9(11):4807–11, Nov 1989.

[8] Jane M Carlton, Robert P Hirt, Joana C Silva, Arthur L Delcher, Michael Schatz, Qi Zhao, Jennifer R Wortman, Shelby L Bidwell, U Cecilia M Alsmark, Sébastien Besteiro, Thomas Sicheritz-Ponten, Christophe J Noel, Joel B Dacks,

Peter G Foster, Cedric Simillion, Yves Van de Peer, Diego Miranda-Saavedra, Geoffrey J Barton, Gareth D Westrop, Sylke Müller, Daniele Dessi, Pier Luigi Fiori, Qinghu Ren, Ian Paulsen, Hanbang Zhang, Felix D Bastida-Corcuera, Augusto Simoes-Barbosa, Mark T Brown, Richard D Hayes, Mandira Mukherjee, Cheryl Y Okumura, Rachel Schneider, Alias J Smith, Stepanka Vanacova, Maria Villalvazo, Brian J Haas, Mihaela Pertea, Tamara V Feldblyum, Terry R Utterback, Chung-Li Shu, Kazutoyo Osoegawa, Pieter J de Jong, Ivan Hrdy, Lenka Horvathova, Zuzana Zubacova, Pavel Dolezal, Shehre-Banoo Malik, John M Logsdon, Katrin Henze, Arti Gupta, Ching C Wang, Rebecca L Dunne, Jacqueline A Upcroft, Peter Upcroft, Owen White, Steven L Salzberg, Petrus Tang, Cheng-Hsun Chiu, Ying-Shiung Lee, T Martin Embley, Graham H Coombs, Jeremy C Mottram, Jan Tachezy, Claire M Fraser-Liggett, and Patricia J Johnson. Draft genome sequence of the sexually transmitted pathogen trichomonas vaginalis. *Science*, 315(5809):207–12, Jan 2007.

[9] Ramu Chenna, Hideaki Sugawara, Tadashi Koike, Rodrigo Lopez, Toby J Gibson, Desmond G Higgins, and Julie D Thompson. Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res*, 31(13):3497–500, Jul 2003.

[10] J M Cherry, C Adler, C Ball, S A Chervitz, S S Dwight, E T Hester, Y Jia, G Juvik, T Roe, M Schroeder, S Weng, and D Botstein. Sgd: Saccharomyces genome database. *Nucleic Acids Res*, 26(1):73–9, Jan 1998.

[11] Peter J A Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J L de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–3, Jun 2009.

[12] Loredana Lo Conte, Steven E Brenner, Tim J P Hubbard, Cyrus Chothia, and Alexey G Murzin. Scop database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res*, 30(1):264–7, Jan 2002.

[13] Inês Cunha-Ferreira, Inês Bento, and Mónica Bettencourt-Dias. From zero to many: control of centriole number in development and disease. *Traffic*, 10(5):482–98, May 2009.

[14] Joel B Dacks, Andrew A Peden, and Mark C Field. Evolution of specificity in the eukaryotic endomembrane system. *Int J Biochem Cell Biol*, 41(2):330–40, Feb 2009.

[15] Petra Fey, Pascale Gaudet, Tomaz Curk, Blaz Zupan, Eric M Just, Siddhartha Basu, Sohel N Merchant, Yulia A Bushmanova, Gad Shaulsky, Warren A Kibbe, and Rex L Chisholm. dictybase–a dictyostelium bioinformatics resource update. *Nucleic Acids Res*, 37(Database issue):D515–9, Jan 2009.

[16] Bianka L Grosshans, Darinel Ortiz, and Peter Novick. Rabs and their effectors: achieving specificity in membrane traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 103(32):11821–7, Aug 2006.

[17] Sarah R Hart, King Wai Lau, Zhiqi Hao, Richard Broadhead, Neil Portman, Andreas Hühmer, Keith Gull, Paul G McKean, Simon J Hubbard, and Simon J Gaskell. Analysis of the trypanosome flagellar proteome using a combined electron transfer/collisionally activated dissociation strategy. *J Am Soc Mass Spectrom*, 20(2):167–75, Feb 2009.

[18] T J P Hubbard, B L Aken, S Ayling, B Ballester, K Beal, E Bragin, S Brent, Y Chen, P Clapham, L Clarke, G Coates, S Fairley, S Fitzgerald, J Fernandez-Banet, L Gordon, S Graf, S Haider, M Hammond, R Holland, K Howe, A Jenkinson, N Johnson, A Kahari, D Keefe, S Keenan, R Kinsella, F Kokocinski, E Kulesha, D Lawson, I Longden, K Megy, P Meidl, B Overduin, A Parker, B Pritchard, D Rios, M Schuster, G Slater, D Smedley, W Spooner, G Spudich, S Trevanion, A Vilella, J Vogel, S White, S Wilder, A Zadissa, E Birney, F Cunningham, V Curwen, R Durbin, X M Fernandez-Suarez, J Herrero, A Kasprzyk, G Proctor, J Smith, S Searle, and P Flicek. Ensembl 2009. *Nucleic Acids Res*, 37(Database issue):D690–7, Jan 2009.

[19] Chandra L Kilburn, Chad G Pearson, Edwin P Romijn, Janet B Meehl, Thomas H Giddings, Brady P Culver, John R Yates, and Mark Winey. New tetrahymena basal body protein components identify basal body domain structure. *J Cell Biol*, 178(6):905–12, Sep 2007.

[20] Kalpana Lal, Mark C Field, Jane M Carlton, Jim Warwicker, and Robert P Hirt. Identification of a very large rab gtpase family in the parasitic protozoan trichomonas vaginalis. *Mol Biochem Parasitol*, 143(2):226–35, Oct 2005.

[21] P Mackiewicz and E Wyroba. Phylogeny and evolution of rab7 and rab9 proteins. *Bmc Evol Biol*, 9(1):101, May 2009.

[22] Aron Marchler-Bauer, John B Anderson, Farideh Chitsaz, Myra K Derbyshire, Carol DeWeese-Scott, Jessica H Fong, Lewis Y Geer, Renata C Geer, Noreen R Gonzales, Marc Gwadz, Siqian He, David I Hurwitz, John D Jackson, Zhaoxi Ke, Christopher J Lanczycki, Cynthia A Liebert, Chunlei Liu, Fu Lu, Shennan

Lu, Gabriele H Marchler, Mikhail Mullokandov, James S Song, Asba Tasneem, Narmada Thanki, Roxanne A Yamashita, Dachuan Zhang, Naigong Zhang, and Stephen H Bryant. Cdd: specific functional annotation with the conserved domain database. *Nucleic Acids Res*, 37(Database issue):D205–10, Jan 2009.

[23] José L Nepomuceno-Silva, Luiz Dione B de Melo, Sergio M Mendonçã, Julio C Paixão, and Ulisses G Lopes. Rjls: a new family of ras-related gtp-binding proteins. *Gene*, 327(2):221–32, Mar 2004.

[24] R Overbeek, M Fonstein, M D'Souza, G D Pusch, and N Maltsev. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6):2896–901, Mar 1999. Bi-directional best hit paper.

[25] Jose B Pereira-Leal. The ypt/rab family and the evolution of trafficking in fungi, Jan 2008.

[26] Jose B Pereira-Leal and M C Seabra. The mammalian rab family of small gtpases: definition of family and subfamily sequence motifs suggests a mechanism for functional specificity in the ras superfamily. *J Mol Biol*, 301(4):1077–87, Aug 2000.

[27] Jose B Pereira-Leal and M C Seabra. Evolution of the rab family of small gtp-binding proteins. *J Mol Biol*, 313(4):889–901, Nov 2001.

[28] Suzanne Pfeffer and Dikran Aivazian. Targeting rab gtpases to distinct membrane compartments. *Nat Rev Mol Cell Biol*, 5(11):886–96, Nov 2004.

[29] Suzanne R Pfeffer. Structural clues to rab gtpase functional diversity. *J Biol Chem*, 280(16):15485–8, Apr 2005.

[30] Emmanuel Quevillon, Tobias Spielmann, Karima Brahimi, Debasish Chattopadhyay, Edouard Yeramian, and Gordon Langsley. The plasmodium falciparum family of rab gtpases. *Gene*, 306:13–25, Mar 2003.

[31] Stephen Rutherford and Ian Moore. The arabidopsis rab gtpase family: another enigma variation. *Curr Opin Plant Biol*, 5(6):518–28, Dec 2002.

[32] Yumiko Saito-Nakano, Brendan J Loftus, Neil Hall, and Tomoyoshi Nozaki. The diversity of rab gtpases in entamoeba histolytica. *Exp Parasitol*, 110(3):244–52, Jul 2005.

[33] A Salminen and P J Novick. A ras-like protein is required for a post-golgi event in yeast secretion. *Cell*, 49(4):527–38, May 1987.

[34] Mami Shintani, Minoru Tada, Tetsuo Kobayashi, Hiroaki Kajiho, Kenji Kontani, and Toshiaki Katada. Characterization of rab45/rasef containing ef-hand domain and a coiled-coil motif as a self-associating gtpase. *Biochem Biophys Res Commun*, 357(3):661–7, Jun 2007.

[35] H Stenmark. Rab gtpases as coordinators of vesicle traffic. *Nat Rev Mol Cell Biol*, Jul 2009.

[36] Van Tompkins, Jussara Hagen, Valerie P Zediak, and Dawn E Quelle. Identification of novel arf binding proteins by two-hybrid screening. *Cell Cycle*, 5(6):641–6, Mar 2006.

[37] P A Wigge, O N Jensen, S Holmes, S Souès, M Mann, and J V Kilmartin. Analysis of the saccharomyces spindle pole by matrix-assisted laser desorption/ionization (maldi) mass spectrometry. *J Cell Biol*, 141(4):967–77, May 1998.

[38] Derek Wilson, Ralph Pethica, Yiduo Zhou, Charles Talbot, Christine Vogel, Martin Madera, Cyrus Chothia, and Julian Gough. Superfamily–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res*, 37(Database issue):D380–6, Jan 2009.

[39] A C Wong, D Shkolny, A Dorman, D Willingham, B A Roe, and H E McDermid. Two novel human rab genes with near identical sequence each map to a telomere-associated region: the subtelomeric region of 22q13.3 and the ancestral telomere band 2q13. *Genomics*, 59(3):326–34, Aug 1999.

[40] Jun Zhang, Karen L Schulze, P Robin Hiesinger, Kaye Suyama, Stream Wang, Matthew Fish, Melih Acar, Roger A Hoskins, Hugo J Bellen, and Matthew P Scott. Thirty-one flavors of drosophila rab proteins. *Genetics*, 176(2):1307–22, Jun 2007.