available at www.sciencedirect.com

SCIENCE DIRECT®

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

A comprehensive comparative analysis of the occurrence of developmental sequences in fungal, plant and animal genomes

*David MOORE**, *Audrius MEŠKAUSKAS*

Faculty of Life Sciences, The University of Manchester, 1.800 Stopford Building, Oxford Road, Manchester M13 9PT, UK

ARTICLE INFO

Article history:
Received 2 September 2005
Received in revised form
2 December 2005
Accepted 19 January 2006
Published online 28 February 2006
Corresponding Editor:
Nicholas P. Money

ABSTRACT

We report a fully comprehensive data-mining exercise, involving an estimated total of 590,000 similarity searches, using agents available on the Internet to search for homologies to polypeptide sequences assigned to the category 'development' in the Gene Ontology Consortium AmiGO database (www.godatabase.org). The results indicate that of 552 such developmental sequences only 78 are shared between all three kingdoms, 72 are shared only between fungi and animals, 58 sequences are shared between plants and fungi, and four sequences were common only to *Dictyostelium* and fungi. No sequences were strictly fungus specific, but 68 occurred only in plants (*Viridiplantae*) and 239 occurred only in animals (*Metazoa*). Although some homology was indicated for a total of 219 fungal sequences, 143 (65 %) of the matches returned were assigned *E*-values of 0.05 and must be categorised as weak similarities at best. The majority of the highly similar matches found in this survey proved to be between sequences involved in basic cell metabolism or essential eukaryotic cell processes (enzymes in common metabolic pathways, transcription regulators, binding proteins, receptors and membrane proteins). What is lacking is cross-kingdom similarity in the management processes that regulate multicellular development. The crown group of eukaryotic kingdoms control and regulate their developmental processes in very different ways. Unfortunately, we know nothing about molecular control of multicellular fungal developmental biology.

Introduction

Current understanding of phylogenetic relationships is that the major kingdoms of eukaryotes separated from one another at a stage before the multicellular grade of organisation. Consequently, there is no logical reason to expect that these kingdoms will share features that contribute to multicellular developmental biology. The fungal hypha differs in so many important respects from animal and plant cells that significant differences in the way cells interact in the construction of organised tissues must be expected (Moore 2005). In the course of their evolutionary history these very different organisms may have needed to solve the same sorts of morphogenetic control problems and may have found some common strategies. Comparison of the way similar functions are controlled can show how different cell biologies can solve similar problems (Meyerowitz 1999), although fungi are not often included in such discussions.

There are now sufficient filamentous fungal genomes in the public sequence databases to make direct sequence comparisons with animal and plant genomes a worthwhile exercise, and a recent search of filamentous fungal genomes with

* Corresponding author. Tel.: +44 1612753903; fax: +44 1612755656.
E-mail address: david.moore@manchester.ac.uk
0953-7562/$ – see front matter
doi:10.1016/j.mycres.2006.01.003

a modest selection of gene sequences generally considered to be essential and highly conserved components of normal development in animals and plants failed to reveal any homologies (Moore *et al.* 2005). This can be taken to suggest that fungal and animal lineages may have diverged from their common opisthokont line (Cavalier-Smith & Chao 1995) well before the emergence of any multicellular arrangement. Presumably, the unique cell biology of filamentous fungi has caused control of multicellular development in fungi to evolve in a radically different fashion from that in animals and plants.

Here we expand the comparison to include all sequences assigned to the biological process 'development'. This is defined as 'biological processes specifically aimed at the progression of an organism over time from an initial condition (e.g. a zygote or a young adult) to a later condition (e.g. a multicellular organism or an aged adult)' in the Gene Ontology Consortium AmiGO (GO) database (www.godatabase.org) (Harris *et al.* 2004). We extend our similarity searches for these developmental sequences to all genomes of cellular organisms included as *Metazoa*, *Fungi*, and *Viridiplantae* in the NCBI Taxonomy database (www.ncbi.nlm.nih.gov/Taxonomy) — reporting on an estimated total of over half a million similarity searches.

Materials and methods

Web agent creation

We use the term 'web agent' in this report for a reusable module that interacts with the Internet seeking the given goal(s), for example, 'get the sequence data', 'get the taxonomy information' or 'get the similarity search results', etc.

We created the agents using an application called Sight, which is a Java™-based package that provides a user-friendly interface to generate and connect agents for automatic genomic data mining (Meškauskas *et al.* 2004). A Sight web agent is essentially an active flow chart in which each element is a working preprogrammed routine. The user assembles the flow chart according to the task s/he wishes to perform. Sight was originally developed to automate analysis of the human genome, but fungus-related Internet resources use different methods of representing the information they report. For the most part, this is a presentational rather than a scientific issue; web pages devoted to such resources were designed later than pages devoted to plant or human sequences. As a result, they make intensive use of advanced features like JavaScript language, unusual (often nested) tables, multiple pages per response, and so on. Such features provide a serious challenge for web agent applications that must still be able to extract a clear data structure from a complicated multi-page server response. The latest version of Sight (version 3.2.0 beta) including features (such as loops and convergences) tailored to servers carrying fungal databases is freely available for download from the project website at http://bioinformatics.org/jSight/.

The Sight web agent executes a single remote or local algorithm and comprises two data structures: one defining the submitted request and the other the received response. The requests contain named values (parameters), required by the Internet service being used. The program uses these request to complete Internet query forms consisting of fields, checkboxes and other controls, just as they are set manually by a human user. In contrast to the request, the result returned to the web agent is often an array of records (for example, several similarities may be found to the search sequence(s), multiple genes in a DNA sequence, multiple motifs in the sequence, etc). Consequently, the Sight agent response is programmed as an array of records that also consist of multiple named fields. As the request and response format differs for each agent, the agents also contain explanatory data structures defining these formats. When the user creates a workflow consisting of several agents, s/he also specifies required assignments between the response parts of the master agent and the request fields of the subsequent slave agent. Default values for request fields may also be included.

Initial query

The initial query in this analysis was to the GO database to get information on genes specifically responsible for a particular biological process; specifically, the progression of an organism over time from an initial condition (e.g. a zygote or a young adult) to a later condition (e.g. a multicellular organism or an aged adult). The vocabulary, of course, is heavily influenced by the fact that by far the majority of entries in genomic databases deal with animal systems, but it is only a vocabulary and does not preclude comparison of the processes in different organisms providing appropriate translation of the terminology can be achieved. In the GO database (Harris *et al.* 2004), the gene group we searched has the accession number GO:0007275 (development), and also belongs to the larger group GO:103163 (biological process). The query was sent to the AmiGO server and the received response stored as a local HTML document.

Retrieving sequences

Each search hit from the AmiGO server contained two hyperlink references, one linking to the entry inside the GO database server itself, and another to the external server from which the original data were derived. Five hundred and fifty-two paired references were returned and this set of pairs was processed by the group of specialised web agents. If the page in the GO database server contained the protein sequence, the sequence was taken from there. Otherwise, the domain of the link to the original data source was checked and one of the specialised sequence retrievers was called. Such retrievers were written for all the web services listed in Table 1.

The number of links to Gramene (Wheeler *et al.* 2003), Rat Genome Database (Twigger *et al.* 2002) and ZFIN were too few to justify writing a tailored web agent, so such sequences were retrieved manually.

Similarity searches

The similarity search (Altschul *et al.* 1990) was performed both in the protein (using BLASTP) and nucleic acid (using TBLAST) sequence databases. Only hits with *E*-value less or equal to 0.05 were accepted (discussed below and see the page dealing with the statistics of sequence similarity scores at: http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html).

Table 1 – Summary of the web services for which web agents were written

Task	Reference	Web server (all services were accessed using the http://communication protocol which is omitted here)
Sequence retrieval	Apweiler *et al.* 2004	www.pir.uniprot.org
Sequence retrieval	Venter *et al.* 1992	www.tigr.org
Sequence retrieval	Huala *et al.* 2001	www.arabidopsis.org
Sequence retrieval	Ashburner & Drysdale 1994	www.flybase.org
Sequence retrieval	Blake *et al.* 2003	www.informatics.jax.org
Sequence retrieval	Kreppel *et al.* 2004	www.dictybase.org
Taxonomy search	Wheeler *et al.* 2005	www.ncbi.nlm.nih.gov/Taxonomy
Similarity search (protein, DNA and RNA)	Venter *et al.* 1992	www.tigrblast.tigr.org
DNA similarity search	These sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/.	www.genome.jgi-psf.org/whiterot1/whiterot1.home.html
Similarity search (protein, DNA and RNA)	Wheeler *et al.* 2005	www.ncbi.nlm.nih.gov/BLAST

Low complexity filters, which remove short, very widespread sequence fragments, were turned on. All other search options were left with the default values proposed by the research groups that administer the search servers.

Each and every 'developmental' sequence retrieved from the AmiGO server was submitted by our web agents to search all the genomes included in the taxonomic listings used by NCBI to specify *Metazoa* (875 genome sequences) and *Viridiplantae* (53 genome sequences) in the Animal and Plant kingdoms, and the entire list included under 'Fungi' (141 genome sequences). An estimated total of 590,000 similarity searches.

Identifying the taxonomic group

The taxonomic group of the organism containing the retrieved sequence was identified in several ways. Some specialised databases contain the sequences either from a single species (like mouse or *Arabidopsis* genomes) or from several species belonging to the same kingdom (like FlyBase), making the taxonomic group evident. Some sequences contained the Latin binomial name of the organism in the sequence header. The taxonomy of such organisms was determined by the web agent by submitting the organism name to the NCBI taxonomy search service (which used the search form at http://www.ncbi.nlm.nih.gov/Taxonomy/) and automatically analysing the web page received in response. For the NCBI nucleic acid database, the search was performed by explicitly limiting the search scope to the given kingdom.

Processing the search data

For each sequence entry from the GO database, we marked the most similar sequence found in an organism from each of the three main eukaryotic groups of multicellular organisms. These data were marked separately for *Metazoa*, *Viridiplantae* and *Fungi* and a final comparison table for all 552 searches was assembled.

The raw data table consisted of all the similarity search results for each sequence from the GO database, together with the BLAST servers involved. The raw data were too bulky and lacking in organisation for reasonable conclusions to be drawn easily and a series of generalisations was performed using an analyzer program operating the following rules:

(i) as the main task was to establish the presence of the sequence in a genome belonging to a certain kingdom, similarity reports from several genomes belonging to that kingdom do not provide more information than a single similarity report; multiple reports were consequently collected and reduced to a single similarity hit.

(ii) to reduce multiple hits in the same kingdom to a single entry in the data table, the hit with the highest similarity score (with repeat filters turned on) was chosen as offering the most reliable conclusions.

(iii) rows of data were then sorted by assigning to each sequence from the GO database a numerical score based on the values of the most significant similarity scores for each of the three kingdoms (*Metazoa*, *Viridiplantae* and *Fungi*) or infinity if hits to a kingdom were not found. In the next step, these values were used to assign a code to each row that identified which kingdoms had returned hits under the chosen critical *E*-value threshold (for instance, *af* for animal + fungi, *apf* for animal + plant + fungi, *ap* for animal + plant, etc). In the final step, rows with identical codes were sorted and counted.

The table of raw data was saved in plain text format as similarity searches were returned, and the easiest way to perform the required generalisations was to write a program to read the raw data, perform the comparisons and sorts, and write another, analyzed, text file. The task is not easy with a database query system, and it proved much simpler to write a specialised Java™ program for this task. This automated analysis greatly reduced the size of the data table by leaving at most three sequences for each search, in cases where similarities were found to be common to all three eukaryotic kingdoms. There were, in fact, a further six similarity combinations: plant–fungal; animal–plant; animal–fungus; plant only; animal only; and similarities found only between *Dictyostelium* sequences and fungal genomes. Thus, seven summary data tables were generated, one for each of these similarity combinations. These summary tables are presented as Supplementary Tables 3–9 at the following URL. The Supplementary

Tables contain descriptive annotations for each sequence, which were retrieved manually from the databases, and live hyperlinks so that users can retrieve the original sequence records directly.

The raw data analyzer is also stored on the Supplementary Data page (source code included), together with the raw data tables (essentially our primary experimental data), all in an archive named 'analyzer.zip'. The multi-database sequence retriever is stored on the same page under the name Sight_polygon.zip. This is the first and the most complicated part of the web agent/robot system used here. It may be a useful example to follow for anyone who wishes to duplicate an analysis of this sort.

However, potential users should be aware that Internet resources change and evolve as database managers and webmasters edit their service pages. For this reason the sequence retriever will need modification, if it is run again, to take account of changes in the design of the websites with which the agent deals. This is why the automated agent generator of the Sight parent application is so important because it permits agents to be tailored to the user's exact requirements and the current structure of the server web pages. However, agents that deal with more complicated workflows involving several servers have the shortest usable lifetimes. In our experience, the Internet services typically remain stable for a month or two, which is enough time to complete the sort of survey reported here. Like building rocket boosters for space missions, a web robot application is built for the mission and used once. For a second mission you need a new robot system; but Sight greatly eases the task of building that system.

Results and discussion

Advantages of web agents over manual searching

The agent algorithm can be written to 'anticipate' a variety of interactions with the search server. For example, they frequently include the ability to follow multiple links, conditional behaviour (for example, an ordered response to a transient server error) and loops (for example, NCBI BLAST can tell the agent to wait for a given duration an arbitrary number of times before returning the results to the agent). So when constructed by an experienced bioinformatician web agents can be at least as effective as a human investigator.

Multiple independent web servers were used in this analysis so it was possible to submit requests in parallel, for which the agents must run in separate execution threads and support task queues. Although this effectively creates a kind of distributed computing, arranging the work of ordinary web servers in parallel significantly differs from distributed computing as this is normally understood. In particular, because each web server is specialised to its own group of tasks, the concept of load balancing is not easily applicable. Also, as we are unable to install the required software on the server(s), the numerous existing tools for distributed computing are not adaptable. However, the possibility of reliably mimicking parallel computing gives the web agents a significant advantage over manual management of this sort of work. While it is certainly possible for an individual worker to submit tasks in parallel from several running instances of their ordinary web browser, frequent switching of concentration between the parallel searches demands extreme skill and minor loss of attention can generate multiple errors. In this the web agents are greatly superior to their human equivalents. Web agents can be arranged in workflows in which they use each other. For example a typical process might be to take the protein sequence from one server and submit it for a similarity search to another server. Then it will be necessary to take part of the header of each returned similarity hit for submission to a third server to identify the taxonomy of the organism. Manual working of this process requires multiple copy/paste operations and switching between several browser windows. Not only does the time required for these clerical operations become comparable with the waiting time for server response, but also while compiling such work a human operator is likely to make mistakes. Web agents can also run around the clock, and use the night hours when servers are less loaded and respond significantly faster. While running, the agent system needs no researcher attention, and the scientist can be busy with other activities.

Humans, of course, can use their knowledge to speed such analysis. For example the Latin binomial name *Homo sapiens* was obviously present in many thousands of the sequence headers that were retrieved. Similarly, *Arabidopsis thaliana* and the names of other popular research organisms were frequently retrieved. A query to the NCBI taxonomy search server typically takes approximately 10 s to execute; a time penalty that is avoided by the human operator's knowledge of the taxonomy. However, this can also be mimicked in the web agents by using a caching system. The agent can then be written to search the cache of previously-retrieved names, a matter of milliseconds only, before issuing a query for any newly-encountered name.

Each web agent system needs researcher time for creating the agents and building the workflow. The time required strongly depends on experience but can be significantly reduced by using specialised development platforms like Sight (Meškauskas *et al.* 2004), as used in this project. Sight has evolved since this initial publication; the latest version providing much better support for agents having multi-step algorithms, including loops and conditional branches (such as agents for using the NCBI BLAST service). We believe that the analysis reported here demonstrates that the web agents generated using such tools can be extremely useful for extensive and highly repetitive tasks. Their role may increase as new methods (Shendure *et al.* 2005) are developed and provide genome sequences for many more species than are currently available.

Comparison of the occurrence of developmental gene sequences in the genomes of eukaryotes

The initial query to the AmiGO server resulted in return of 552 sequences categorised as developmental, of which only 78 are shared between all three kingdoms, 72 are shared only between fungi and animals, 58 sequences are shared between plants and fungi, and four sequences were common only to *Dictyostelium* and fungi (Table 2).

No sequences were strictly fungus specific, but 68 occurred only in *Viridiplantae* and 239 occurred only in *Metazoa*. In many respects these constitute 'control' searches in that they

Table 2 – Summary of significant similarities returned

Kingdom	Hits	Remarks
Animal only	239	All had *E*-values well below 0.05
Plant only	68	All had *E*-values well below 0.05
Fungi and *Dictyostelium* only	4	Of which three had *E*-values of 0.05 (and the fourth an *E*-values of 0.03)
Animal and plant	33	Of which 13 had *E*-values of 0.05
Fungi and animal	72	Of which 64 had *E*-values of 0.05
Fungi and plant	58	Of which 55 had *E*-values of 0.05
Common to all three kingdoms	78	Of which 14 plant homologies had *E*-values of 0.05, and 20 fungal homologies had *E*-values of 0.05
Total	552	Of which 219 showed some homology with fungal sequences, though 143 of these had *E*-values of 0.05

E-values indicate the likelihood of that similarity between the sequences being found by chance. *E*-values less than 0.01 are numerically very similar to probability statements. So *E*-values of 0.05 mean that there's more than one chance in 20 of the similarity being found by chance — and little significance is assigned to these (even if not due entirely to chance they are likely to indicate possession of similar functional motifs — e.g. shared DNA binding sites, membrane spanning regions, etc.).

represent successful returns within the kingdom from which the original reference sequence was obtained. It is significant, therefore, that all of these similarities have *E*-values markedly less than the arbitrary cut-off value of 0.05. This validates the process by showing that the reference sequences can be shown to retrieve highly similar sequences with high probability from within their own kingdom. Unfortunately, there are no fungal sequences that are categorised as being involved in developmental processes. This is not a fault in the GO database; rather it is a deficiency that justly reflects the low level of research interest in the developmental biology of kingdom *Fungi*.

Although in the case of Kingdom *Fungi* they are mostly negative, cross-kingdom comparisons are interesting. Searches with 44 'no apical meristem' (NAM) family proteins failed to detect any similarities with animals or fungi (Supplementary Table 3); a further 42 NAM family protein sequences showed weak similarities (*E*-value = 0.05) with fungal genomes, but still with no similarity in *Metazoa* (Supplementary Table 4 — Plant–fungal homologies). The one exception in this protein family is NAM locus AT4G28500 (a predicted protein of *Arabidopsis thaliana* with transcription factor activity) for which the search revealed homology (*E*-value = 5×10^{-5}) with clone RP11-26F2 of the *Homo sapiens* chromosome 15, and a weak similarity (*E*-value = 0.05) with the fungal *Phanerochaete* genome (which is not annotated; Supplementary Table 5). This suggests that the developmental functions represented by the family of NAM proteins are restricted to plants.

The 'seven in absentia' (SINA) protein family, which is required for photoreceptor cell formation during *Drosophila* eye development, has a more mixed distribution. No similarity in metazoan or fungal genomes can be detected for eight of the *Arabidopsis* SINA proteins (Supplementary Table 3), but high levels of similarity (*E*-values less than 10^{-28}) emerged for four other *Arabidopsis* SINA proteins (Supplementary Table 6), and the mouse siah2 protein showed moderate homology with a protein from the green alga *Spermatozopsis* (*E*-value = 9.6×10^{-3}) and a predicted protein of *Neurospora crassa* (*E*-value = 5.6×10^{-3}; Supplementary Table 5). In view of the suggestion that SINA proteins may mediate p53-dependent cell-cycle arrest in man (Matsuzawa *et al.* 1998), it is interesting that 16 mammalian p53 protein sequences (Supplementary Tables 5 and 6) showed very high similarity (*E*-values less than 10^{-100}) with a protein from the *Zea mays* genome. Oddly, eight of the mammalian p53 sequences failed to detect similarity with fungal genomes (Supplementary Table 6), although sequences from African green monkey, Chinese hamster, rhesus monkey, tree shrew and woodchuck all showed complete homology (*E*-value = 0) with a predicted mRNA reported from the *Ustilago maydis* genome, whilst gerbil, porcine and guinea pig sequences were weakly similar (*E*-value reported as <0.05) to sequences in the *Phanerochaete* genome (Supplementary Table 5).

Only three plant sequences retrieved highly similar sequences from the fungal genomes (Supplementary Table 4). These were the phosphoribosylanthranilate isomerase of *Arabidopsis thaliana*, which is similar to the TRP-F sequence of *Candida glabrata* (*E*-value = 1.2×10^{-21}); a putative oxidoreductase of *Arabidopsis* that is highly similar to a putative dehydrogenase/reductase of *Aspergillus fumigatus* (*E*-value = 7×10^{-18}); and a hypothetical protein of *Arabidopsis* highly similar (with an *E*-value = 6.3×10^{-10}) to the mybC transcription factor of *Dictyostelium* and to a hypothetical protein of *Candida albicans* (*E*-value = 4.8×10^{-10}). All other plant–fungus similarities were returned with *E*-value reported as 0.05 (weak similarity).

In the list of animal–plant similarities (Supplementary Table 6), apart from the SINA and p53 similarities already noted, very low *E*-value similarities were limited to two sialyltransferases (*E*-value reported as zero), a cytosine methyl transferase (*E*-value = 4.8×10^{-9}), a transcription factor (*E*-value = 9.9×10^{-16}), a transcriptional co-activator (*E*-value = 2.5×10^{-3}), a receptor protein (*E*-value = 5.6×10^{-3}) and a homeobox domain protein (*E*-value = 2.8×10^{-2}). All other plant–animal similarities were returned with *E*-values of 0.05 (Supplementary Table 6).

Most animal–fungus similarities were also weak (Supplementary Table 7). A predicted mRNA from the *Ustilago maydis* genome proved to be homologous (*E*-value = 0) to the ISL1 mouse transcription factor, and a hypothetical protein of *U. maydis* was very similar (*E*-value = 1.6×10^{-5}) to the human orthologue of the *pad*-1 gene of *Caenorhabditis elegans*, which is required for embryonic patterning. *E*-values in the region of 10^{-3} were returned to a Zebrafish nuclear respiratory factor (with a potential cell surface flocculin of *Candida albicans*), a *Dictyostelium* actin binding protein (with a hypothetical protein of *Magnaporthe grisea*), a human ATP-dependent DNA helicase [with a hypothetical protein from *Eremothecium* (*Ashbya*) *gossypii*], and a human Ariadne-2 protein homolog [with a hypothetical protein of *Gibberella zeae* (anam. *Fusarium graminearum*)]. All other fungus–animal similarities were returned with *E*-values of 0.05 (Supplementary Table 7). Much the same applies to the four *Dictyostelium* sequences, which failed to retrieve any similarities in either *Metazoa* or *Viridiplantae*, but were detectable in fungal genomes (Supplementary Table 8). One, a putative GATA-binding transcription factor of *Dictyostelium* was marginally similar to a hypothetical protein of *Gibberella zeae* (*E*-value = 2.8×10^{-2}), but the other three (two transcription regulators and an adhesion modulator) returned similarities in *Cryptococcus* and *Phanerochaete* with *E*-values of 0.05.

Although some homology was indicated for a total of 219 sequences from fungal genomes, 143 (65 %) of the matches returned were assigned *E*-values of 0.05. This level of similarity corresponds approximately to a probability of 1 in fewer than 20 of finding the match purely by chance, and we believe this to be too low a level of similarity for much significance to be assigned to it.

The majority of the highly similar matches found in this survey proved to be between sequences involved in what could be described as basic cell metabolism or essential eukaryotic cell processes. Within this group are found enzymes in common metabolic pathways, many transcription regulators, binding proteins, receptors and membrane proteins. What is lacking is cross-kingdom similarity in the 'higher-management' functions that integrate these 'nuts and bolts' of development. In particular, it is evident that NAM sequences are essentially limited to plants. We have previously made a search of filamentous fungal genomes with a modest selection of gene sequences generally considered to be essential and highly conserved components of normal development in animals and plants, which failed to reveal any homologies (Moore *et al.* 2005), from which we concluded that the major eukaryotic lineages diverged well before the emergence of any multicellular arrangement. Such a conclusion seems to be amply supported by this more comprehensive survey, which shows that three Notch, four TGF-β, and 13 Wnt sequences (all widely considered as essential, highly conserved, components of normal development in animals) fail to retrieve sequences showing any significant similarity from any of the plant or fungal genomes currently available (Supplementary Table 9). A weak similarity was detected between Zebrafish Wnt-4a protein precursor and a putative ubiquitin-specific protease 1 of *Arabidopsis* (Supplementary Table 6), but was assigned an *E*-value of 0.05 and its significance must await detailed comparison of the sequences.

Overall, our findings suggest that there is no strong resemblance between the crown group of eukaryotic kingdoms in the way they control and regulate their developmental processes. Perhaps it is time for some real effort to be made to find out how the members of kingdom *Fungi* manage their multicellular morphogenesis.

Supplementary material

Supplementary material associated with the article can be found, in the online version at doi: 10.1016/j.mycres.2006.01.003.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS, 2004. UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Research* **32**: D115–D119.

Ashburner M, Drysdale R, 1994. FlyBase — the *Drosophila* genetic database. *Development* **120**: 2077–2079.

Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT, 2003. MGD: the Mouse Genome Database. *Nucleic Acids Research* **31**: 193–195.

Cavalier-Smith T, Chao EE, 1995. The opalozoan *Apusomonas* is related to the common ancestor of animals, fungi and choanoflagellates. *Proceedings of the Royal Society of London Series B* **261**: 1–6.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la CN, Tonellato P, Jaiswal P, Seigfried T, White R, 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* **32**: D258–D261.

Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W, Mueller LA, Bhattacharyya D, Bhaya D, Sobral BW, Beavis W, Meinke DW, Town CD, Somerville C, Rhee SY, 2001. The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Research* **29**: 102–105.

Kreppel L, Fey P, Gaudet P, Just E, Kibbe WA, Chisholm RL, Kimmel AR, 2004. DictyBase: a new *Dictyostelium discoideum* genome database. *Nucleic Acids Research* **32**: D332–D333.

Matsuzawa S, Takayama S, Froesch BA, Zapata JM, Reed JC, 1998. p53-inducible human homologue of Drosophila seven in absentia (Siah) inhibits cell growth: suppression by BAG-1. *EMBO Journal* **17**: 2736–2747.

Meškauskas A, Lehmann-Horn F, Jurkat-Rott K, 2004. Sight: automating genomic data-mining without programming skills. *Bioinformatics* **20**: 1718–1720.

Meyerowitz EM, 1999. Plants, animals and the logic of development. *Trends in Cell Biology* **9**: M65–M68.

Moore D, 2005. Principles of mushroom developmental biology. *International Journal of Medicinal Mushrooms* **7**: 79–102.

Moore D, Walsh C, Robson GD, 2005. A search for developmental gene sequences in the genomes of filamentous fungi. In: Arora DK & Berka R, (eds), *Applied Mycology and Biotechnology, vol. 5, Genes and Genomics*. Elsevier Science, Amsterdam, pp. 169–188.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM, 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**: 1728–1732.

Twigger S, Lu J, Shimoyama M, Chen D, Pasko D, Long H, Ginster J, Chen CF, Nigam R, Kwitek A, Eppig J, Maltais L, Maglott D, Schuler G, Jacob H, Tonellato PJ, 2002. Rat Genome Database (RGD): mapping disease onto the genome. *Nucleic Acids Research* **30**: 125–128.

Venter JC, Adams MD, Martin-Gallardo A, McCombie WR, Fields C, 1992. Genome sequence analysis: scientific objectives and practical strategies. *Trends in Biotechnology* **10**: 8–11.

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L, 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Research* **31**: 28–33.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E, 2005. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **33**: D39–D45.