Pacific Symposium on Biocomputing 6:384-395 (2001)

TEXTQUEST: DOCUMENT CLUSTERING OF MEDLINE ABSTRACTS FOR CONCEPT DISCOVERY IN MOLECULAR BIOLOGY

I. ILIOPOULOS, A. J. ENRIGHT, C. A. OUZOUNIS

Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK

{ioannis, anton, christos}@ebi.ac.uk

We present an algorithm for large-scale document clustering of biological text, obtained from Medline abstracts. The algorithm is based on statistical treatment of terms, stemming, the idea of a 'go-list', unsupervised machine learning and graph layout optimization. The method is flexible and robust, controlled by a small number of parameter values. Experiments show that the resulting document clusters are meaningful as assessed by cluster-specific terms. Despite the statistical nature of the approach, with minimal semantic analysis, the terms provide a shallow description of the document corpus and support concept discovery.

1. Introduction

The vast accumulation of electronically available textual information has raised new challenges for information retrieval technology. The problem of *content analysis* was first introduced in the late 60's [1]. Since then, a number of approaches have emerged in order to exploit free-text information from a variety of sources [2, 3].

In the fields of Biology and Medicine, abstracts are collected and maintained in Medline, a project supported by the U.S. National Library of Medicine (NLM)[1]. Medline constitutes a valuable resource that allows scientists to retrieve articles of interest, based on keyword searches. This query-based *information retrieval* is extremely useful but it only allows a limited exploitation of the knowledge available in biological abstracts.

Query-based retrieval is useful for content-focused querying [4], where searches pre-suppose that the end-user is familiar with the subject at hand or that they would know precisely what keywords should be used to search for particular items. This, however, is rarely the case, especially in rapidly changing fields such as molecular biology and medicine, where subjects can be extremely complex: there are many synonym terms, new connections are constantly discovered between previously unrelated subjects and review articles are outdated very quickly. In these situations, query-based retrieval usually results in many hours of confusing and misleading searches.

[1] http://www.nlm.nih.gov/

Put in another way, the traditional query-based information retrieval from Medline is less useful when specific and substantial information is necessary in a very short period of time and without the need to read through all the retrieved articles, extracted from a series of searches. In order to exploit successfully this vast amount of textual information, there is an imperative need to find elegant and accurate ways of extracting the desirable biological knowledge.

At the other extreme of the spectrum in text processing, *information extraction* allows the most sophisticated use of syntactic and semantic analysis, providing extensive 'understanding' of free text by statistical and algorithmic approaches [5]. However, these procedures usually require complex software systems, with low precision and recall, a predefined ontology for the domain of discourse and various metrics that may be limited in scope and scale.

An intermediate stage of text analysis involves *document clustering*, where similar documents are detected on the basis of their sharing of particular, meaningful terms [6]. The problem can be stated as follows: given an arbitrary set of documents X, return an optimal set of clusters Y that contain these documents, plus the features Z_{Yi} for each document i in cluster Y. The advantages of document clustering include the small set of heuristics employed, a minimum amount of semantic analysis and its reliance on simple statistical measures.

This generic approach to extract knowledge that is embedded in textual information is provided by *unsupervised machine learning* [7]. In this approach, a set of instances containing a number of features are fed into clustering algorithms that generate optimal classification schemes which maximally describe the data, in terms of clusters, decision trees and the like. In this context, the unsupervised machine learning approach has also been defined as '*concept discovery*' [8] or (the much harder problem of) '*ontology induction*', where classifications are automatically derived from vast amounts of data.

There has been a growing interest in biological text processing during the past few years [9], mostly focusing on two general areas: the detection and extraction of relations [10-12] and the detection of keywords [13-15]. Herein, we approach the problem of biological text processing from a different perspective, by applying document clustering using term co-occurrence. The procedure uses terms to associate documents, while the end result is also the detection of the most significant keywords that yield the document clusters. From this perspective, our approach relates mostly to keyword extraction although relations are also recorded in the form of term co-occurrence. The only previous attempt for document clustering of biological text has been developed by NLM for the 'neighbors' utility [16].

This approach has wider implications, because of its general applicability to a number of text analysis problems. Currently, one of the central problems in bioinformatics is the issue of data retrieval and integration. Despite an influx of molecular data in the form of sequences, structures, transcription profiles etc., the

real body of biological knowledge comes in the form of abstracts, and soon in the form of publication repositories, such as PubMed Central[2] and BioMed Central[3]. Medline abstracts provide a basis for experimentation with text analysis, in a highly complex, heterogeneous and constantly changing information landscape. The challenge for bioinformatics is to transform and integrate automatically and reliably large volumes of both molecular and textual information, to provide prototypes for the next-generation of database systems that will support biological research. The ultimate criterion for the performance of these systems is whether they generate desirable information that is not obtainable otherwise.

2. Methods

In our method presented here, Medline abstracts are selected, processed through a successive number of steps and re-structured to obtain the optimal number of terms that would associate large numbers of biological documents into some coherent and meaningful groups, potentially representing the biological role of particular molecules and processes.

We have devised the following ten-step protocol:

1. A set of abstracts K is selected from Medline using some informed query-based information retrieval. Keywords that are pertinent to a biological process or a single species are used to focus subsequent processing on a specific document set. The set is saved in 'Medline' format.
2. The selected abstracts are parsed and only the fields 'UID' (unique Medline identifier) and 'AB' (abstract body) are retained. The MeSH terms are not used, because we have found that they are not always up-to-date and may not reflect the contents of particular document sets. The set of terms S at this step is quite large and redundant, because many documents contain similar terms.
3. To eliminate common English words, one commonly used strategy in text analysis is the generation of a so-called 'stop-list' [17]. Due to the complex character of the document sets returned by arbitrary queries (from step 1), it was not possible to produce a well-defined stop-list. Instead, we have employed the TF.IDF family of metrics, a well-known term weighting scheme [4, 18]. We have used the following variant: $w_i = \log_2(N_i/n_i)$, where w_i is the weight of term i in the document, N_i is the frequency of term i in a reference set L, and n_i is the number of documents in L that term i occurs in. The reference set used was the British National Corpus (BNC)

[2] http://pubmedcentral.nih.gov/
[3] http://www.biomedcentral.com

collection[4]. Terms that appear frequently in a document (TF = Term Frequency), but rarely in the reference set (IDF = Inverse Document Frequency) are more likely to be specific to the document. At this stage, all non-alphanumeric symbols are deleted and all characters are converted into low case: terms may thus be composed by more than one English words.

4. Terms with a high TF.IDF value (typically more than 15) or absent from the BNC collection (for which TF.IDF is not applicable) are retained for further processing. Common English words are thus eliminated. The cut-off value ρ was decided after extensive experimentation and evaluation with various datasets. The set of terms T at this step is also redundant, but reduced considerably, depending on the TF.IDF threshold.
5. All terms from a set of documents (abstracts) are then combined, their frequencies in the query set are counted and the most frequently occurring unique terms are retained. The cut-off threshold σ here is typically 1% of the document number K, e.g. for 1,000 abstracts, *unique* terms that occur at least 10 times are kept for further processing. This step ensures that infrequent terms are eliminated, even if they have a high TF.IDF score, to facilitate clustering. We define the resulting set of terms U as a 'go-list'. This set contains a non-redundant set of terms that are present in the set of abstracts K and satisfy the above mentioned criteria.
6. The go-list is subjected to a simple stemming procedure. This operation eliminates suffixes from related terms and results in a non-redundant list of terms [1]. Stemming is performed as follows: terms that are less than five characters long are kept intact (short terms usually represent gene or protein names). Longer terms are compared to each other with a simple string match: when they share a common string, their two last characters are eliminated. Thus, the resulting set of terms V contains terms that share a common root. Synonym weighting [1] was not used, because of the unavailability of a number of synonyms especially for genes and proteins, as well as a continuing influx of new biological terms.
7. The stemmed go-list is then combined with the set of abstracts to generate a file suitable for the subsequent clustering step. Each abstract is represented by a simple bit vector Γ, where the presence or absence of a term in the go-list is marked by a 1 or 0, respectively. The length of the vector Γ is equal to the size of the stemmed go-list V. Typically, the length of Γ is around 300 bits (representing terms), depending on the size of K (number of abstracts). Notice that the original set of terms S is reduced significantly at this stage (see Table 1 for an example).

[4] ftp://ftp.itri.bton.ac.uk/pub/bnc/all.num.o5

8. The fixed-length array bit vector representation of the abstracts (and their term contents) is then used as input for the unsupervised machine learning step. We have used SGI's MineSet™ data mining software[5] on a 2-processor SGI Octane workstation with 256 MB of memory. We deployed the well-known k-means unsupervised clustering algorithm using a variety of parameter values and multiple iterations. The visualization engine of MineSet™ was also used to interact with the data and the results. Terms that are found to co-occur in a large number of abstracts provide the signal for the machine learning algorithm to discover groups of related abstracts. Using terms as characters for the clustering procedure not only generates sets of related abstracts but the very same terms actually characterize the content of the obtained clusters. Thus, the method is a combination of document clustering and concept discovery. Evidently, other machine learning methods are possible at this step.
9. To obtain the final set of terms W that are specific and highly descriptive for a given cluster of documents, we employ the well-known log-odds formula: $\theta_{ij} = \log_2(f_{ij}/f_i)$, where θ_{ij} represents the preference of term i in a document cluster j, f_{ij} represents the frequency of term i in cluster j and f_i represents the frequency of term i in the total set of abstracts. If $f_{ij} = f_i$ (i.e. term i is as frequently found in a cluster j as in the total set), then θ_{ij} is zero. Positive values of θ characterize terms specific to a cluster - and vice versa. Usually, we obtain terms with a positive θ value, greater than a cut-off threshold value τ.
10. Results are visualized using `xlayout` (Enright, unpublished), an optimization and graphical display method based on graph drawing by the force-directed placement algorithm [19, 20].

The above protocol is based on the statistical treatment of words, a minimal amount of modifications (e.g. stemming) and clustering using unsupervised learning. All the threshold values (ρ, σ, τ) that have been used were optimized empirically, by extensive experimentation, and can be set as parameters by the user. It is worth noting that the 'semantics' of text have been encoded in these various steps in the form of heuristics (e.g. short terms that are significant usually represent gene names, thus they are not stemmed). We have not attempted to use any explicit form of word disambiguation or devise an ontology for a specific domain. Despite that, the performance of the system is surprisingly high (see Results). Our approach was to keep the architecture of the system as general as possible, without encoding facts pertinent to a specific biological process.

The precision of the system (words that are not particularly meaningful but are present in the final clusters) is quite impressive (see Results). Recall cannot be

[5] http://www.sgi.com/software/mineset/

estimated, because there is a lack of test sets for problems of this size. We, instead, relied on extensive evaluation of various experiments. We have applied our method to numerous datasets, and the results obtained were checked manually for consistency. As mentioned above, the terms of the stemmed go-list play a dual role both as features for unsupervised learning and as descriptors for the contents of the resulting clusters.

3. Results

To evaluate the reliability of our approach, we have performed various control experiments, one of which is presented here in some detail. We have obtained an equal number of 830 abstracts from two keyword-based queries: "(escherichia AND pili)" and "(cerevisiae AND cdc*)", accessing all available fields in Medline. We then merged the two sets into one, comprising 1660 abstracts. The clustering should produce two, rather distinct, clusters, with the most significant terms standing out, describing the two sets of abstracts with meaningful keywords. The term reduction process is shown in Table 1 (see Methods for details).

Set of terms	Step	Number of terms
S	2	162,499
T	4	56,057
U	5	868
V	6	633
W	9	471
$\rho = 17$	$\sigma = 0.01$ K	$\tau = 0.8$

Table 1. Term reduction for the control experiment described in text.
Set of terms, Step and parameters (ρ, σ, τ) are described in the Methods section.

The iterative k-means clustering produces indeed two very distinct clusters, with descriptive terms for the two rather different model species and the associated processes, captured by our initial keyword search. Characteristic terms with high log-odds θ values are shown in Table 2 (cut-off τ value was 0.8). It is interesting to note that many terms describe the clusters with high fidelity, and immediately imply the nature of the cluster contents. Unsatisfactory terms (in italics) usually refer to various species (e.g. "melanogaster", "shigella") or experimental techniques (e.g. "elisa", "precipitation"). On the other hand, terms such as "cln1" or "rad9" refer to gene/protein names involved in these processes quite accurately. Also species-specific terms are detected (e.g. "nucleus" for yeast and "operons" for *Escherichia*

coli). Combined together, the collection of terms in this experiment provides some shallow description of the cluster under investigation (Table 2).

Cluster 1	**Cluster 2**
alphafactor	adherence
budding	agglutination
centromere	antigenic
chromatin	bacteriophage
cln1	*chloroform*
cytoskeleton	conjugative
defines	diarrhoea
diploid	*elisa*
fission	fimbrial
gtpbinding	glycoproteins
meiosis	*klebsiella*
melanogaster	*morphologically*
microtubules	operons
nucleus	pfimbriae
phosphorylation	plasmidencoded
rad9	*precipitation*
rescues	pyelonephritis
spindle	serogroup
telomere	*shigella*
tumor	*susceptible*
ubiquitin	uropathogenic
uv	vaccination

Table 2. Representative terms describing two clusters in a control experiment. For details, see text. Italics signify unsatisfactory terms. Notice that hyphens and other non-alphanumeric characters have been deleted.

In a more adventurous experiment, we have selected 525 abstracts from Medline by combining an exhaustive set of abstracts containing the following terms: “anterior-posterior AND drosophila” plus “dorsal-ventral AND drosophila”. Our go-list contained 409 terms. The parameter values for this experiment were: $\rho = 19$, $\sigma = 0.01K$ and $\tau = 0.5$. Our clustering procedure consistently returned three clusters.

Each cluster refers to a set of abstracts that are related by terms that co-occur among the different abstracts. These terms enable us to 'label' each cluster.

The first cluster consists of 206 abstracts and contains names of genes (e.g. antp, bithorax, ftz, wg, hunchback, distalless, engrailed, ubx, eve etc.) and other terms related to the development of *Drosophila* (e.g. segmental, homeobox, homeodomain, stripe, proximal-distal, parasegment, blastoderm etc.) This group of abstracts is related to the process of segmentation and embryonic patterning (reviewed in [21-23]). The most significant terms that describe the cluster successfully refer to gene or protein domain names (e.g. homeobox) or keywords (e.g. segmental) associated with this particular process.

The second cluster is slightly larger, containing 251 abstracts. The gene names (e.g. pelle, notch, cactus, tld, dpp, rel, serrate etc.) and the terms (e.g. dorsalizing, ventralspecific, gastrulation etc.) reveal that the abstracts in this cluster are related to the embryonic dorsoventral axis specification in *Drosophila* (reviewed in [24, 25]).

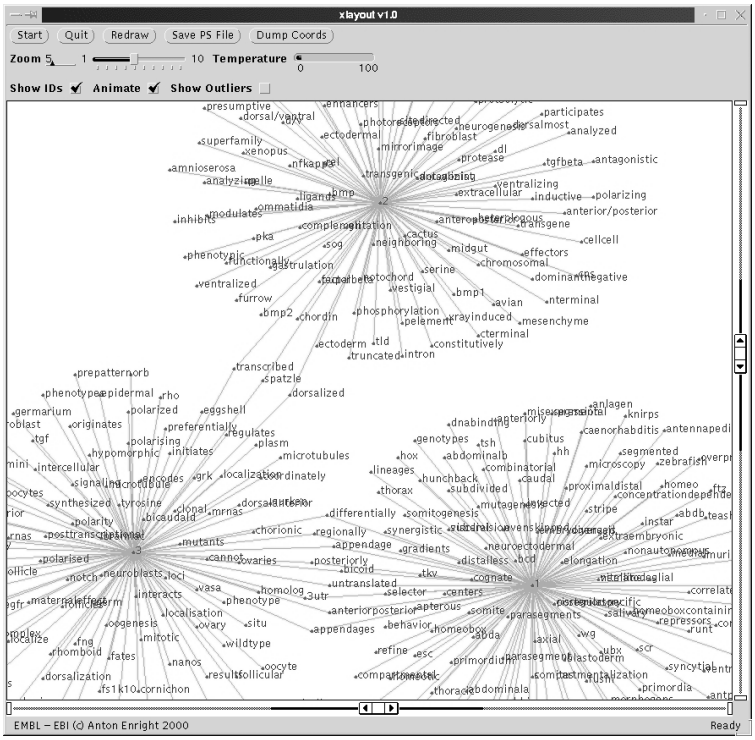

Figure 1. Clusters of terms referring to *Drosophila* embryonic development. Output produced by `xlayout` . Edges (lines) of the graph signify membership to a cluster –length does not reflect any weighting scheme.

The number of abstracts in the third cluster is 68. This group of abstracts can be labeled as 'egg chamber/oocyte patterning', as deduced by the gene names (nanos, oskar, grk, torpedo, vasa, tgf etc.) and the terms (oocyte, maternal-effect, germline, polarized, oogenesis etc.) that belong to this cluster (reviewed in [26, 27]). It is interesting to note that although the initial set of abstracts was obtained with two queries, the resulting clustering procedure using term co-occurrence produces a third cluster, which refers to genes that are involved in the polarization of both the anterior-posterior and the dorsal-ventral axes during *Drosophila* oogenesis [26-28]. Thus, the third cluster represents terms that describe the early developmental stage of the processes represented by the other two, otherwise unrelated, document clusters.

In summary, we have asked the question how are the two *Drosophila* embryonic axes established by choosing 'anterior-posterior' and 'dorsal-ventral' using relevant keywords. The result of this experiment is the automatic clustering of documents, associated with descriptive terms, presented in a concise, rapid and highly non-trivial fashion (Figure 1).

4. Discussion

It appears that term co-occurrence and the processing steps that we have implemented generate reliable document clusters that not only associate Medline abstracts into meaningful groups but also provide the 'labels' for a crude content analysis in a rapid and reliable fashion.

The method is sufficiently flexible and parameter-based to allow the extensive exploration of various document collections. Our data set is available on our group's web site[6]. By plugging in a different parser at step 2, various textual resources can also be analyzed. One obvious example in the field of bioinformatics is the annotation corpus of various sequence databases, such as SwissProt[7].

The performance of the method is influenced by a number of factors. The CPU requirements for the experiments described above are of the order of 1-2 hours. The memory requirements are quite excessive, and there is a trade-off between the number of abstracts (instances) K and the number of terms (features) W. The feature/instance ratio may have to be reduced for very large-scale experiments.

The precision of the method appears to be high, as judged by the relatively few terms that appear to be irrelevant in our experiments. However, accuracy metrics such as precision and recall are not critical in document clustering, as long as the

[6] http://www.ebi.ac.uk/research/cgg/mining/textquest/
[7] http://www.expasy.ch/sprot/

results are meaningful and provide intelligent guidance to the end-user, e.g. for more elaborate queries or a quick summary of vast amounts of literature.

With sufficient computational resources, the method can be applied to 10-100,000 abstracts, possibly describing a particular biological process or the biology of a single species. For example, we envision applying document clustering to all abstracts that refer to *Drosophila* development and obtain automatically an atlas of the major groups of genes and their actions that influence the ontogeny of this model organism.

It may very well be that there is a set of terms that have been used specifically for certain biological species, while other terms are shared across related species. In analogy with the terms 'genome' or 'proteome', we can define the set of terms that refer to a particular species as the 'conceptome' of that species. This set of terms may be used for comparative studies similar in spirit to comparative genomics [29].

We are in the process of scaling up our computations to perform very large-scale experiments to test some of the above conjectures. We are particularly interested in ontology induction experiments for the automatic classification of proteins into functional classes [30], the discovery of new functional relationships in protein families and data integrity checks of database annotations. The method can also be used for query relaxation in molecular biology databases and massive annotation of large-scale experiments, e.g. transcription profiling.

Acknowledgments

This work was supported by the European Molecular Biology Laboratory and the European Commission (DG-XII – Science, Research and Development). I.I. is a recipient of a TMR Postdoctoral Fellowship. Earlier contributions by Mark Carroll and discussions with members of the Computational Genomics Group are greatly acknowledged. C.O. thanks the UK Medical Research Council and IBM Research for further support.

References

1. G. Salton, "Automatic content analysis in information retrieval" (University of Pennsylvania, PA, 1968)
2. G. Salton, "Automatic Text Processing: the transformation, analysis, and retrieval of information by computer" (Addison-Wesley, Reading MA, 1989)
3. G. Salton, "Developments in automatic text retrieval" *Science* **253**, 974 (1991)

4. Y. Yang, J.G. Carbonel, R.D. Brown, T. Pierce, B.T. Archibald and X. Liu, "Learning approaches for detecting and tracking news events" *IEEE Intelligent Systems* 32 (1999)
5. J. Allen, "Natural language understanding" (Benjamin Cummings, Redwood City CA, 1994)
6. R. Willett, "Recent trends in hierarchic document clustering: a critical review" *Information Processing & Management* **25**, 577 (1988)
7. M. Kubat, I. Bratko and R.S. Michalski in *Machine Learning and Data Mining: methods and applications*, "A review of machine learning methods" Ed. R.S. Michalski, I. Bratko and M. Kubat - p. 3 (John Wiley & Sons, New York NY, 1997)
8. J. Gennari, P. Langley and D. Fisher, "Models of incremental concept formation" *Artificial Intelligence* **40**, 11 (1989)
9. M.A. Andrade and P. Bork, "Automated extraction of information for molecular biology" *FEBS Lett.* **476**, 12 (2000)
10. C. Blaschke, M.A. Andrade, C. Ouzounis and A. Valencia, "Automatic extraction of biological information from scientific text: protein-protein interactions" Intelligent Systems for Molecular Biology, Heidelberg - p. 60 (1999)
11. T. Sekimizu, H.S. Park and J. Tsujii, "Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts" Genome Informatics Workshop, Tokyo - p. 62 (1998)
12. J. Thomas, D. Milward, C. Ouzounis, S. Pulman and M. Carroll, "Automatic extraction of protein interactions from scientific abstracts" *Pac. Symp. Biocomput.* 538 (2000)
13. M.A. Andrade and A. Valencia, "Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families" *Bioinformatics* **14**, 600 (1998)
14. D. Proux, F. Rechenmann, L. Julliard, V. Pillet and B. Jacq, "Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction" Genome Informatics Workshop, Tokyo - p. 72 (1998)
15. K. Fukuda, A. Tamura, T. Tsunoda and T. Takagi, "Toward information extraction: identifying protein names from biological papers" *Pac. Symp. Biocomput.* 707 (1998)
16. W.J. Wilbur and Y. Yang, "An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts" *Comput. Biol. Med.* **26**, 209 (1996)
17. E.M. Voorhees in *WordNet: an electronic lexical database*, "Using WordNet for text retrieval" Ed. C. Fellbaum - p. 285 (MIT Press, Cambridge MA, 1998)
18. J. Rocchio in *The SMART retrieval system - experiments in automated document processing*, "Relevance feedback information retrieval" Ed. G. Salton - p. 313 (Prentice-Hall, Englewood Cliffs NJ, 1971)

19. P. Eades, "A heuristic for graph drawing" *Congressus Numerantium* **42**, 149 (1984)
20. T.M.J. Fruchterman and E.M. Reingold, "Graph drawing by force-directed placement" *Software - practice and experience* in press (2000)
21. L. Pick, "Segmentation: painting stripes from flies to vertebrates" *Dev. Genet.* **23**, 1 (1998)
22. M.D. Biggin and W. McGinnis, "Regulation of segmentation and segmental identity by *Drosophila* homeoproteins: the role of DNA binding in functional activity and specificity" *Development* **124**, 4425 (1997)
23. M. Mannervik, "Target genes of homeodomain proteins" *BioEssays* **4**, 267 (1999)
24. R. Steward and S. Govind, "Dorsal-ventral polarity in the *Drosophila* embryo" *Curr. Opin. Genet. Dev.* **3**, 556 (1993)
25. M.P. Belvin and K.V. Anderson, "A conserved signaling pathway: the *Drosophila* toll-dorsal pathway" *Ann. Rev. Cell Dev. Biol.* **12**, 393 (1996)
26. R.P. Ray and T. Schupbach, "Intercellular signaling and the polarization of body axes during *Drosophila* oogenesis" *Genes Dev.* **10**, 1711 (1996)
27. C. Rongo, H.T. Broihier, L. Moore, M. van Doren, A. Forbes and R. Lehmann, "Germ plasm assembly and germ cell migration in *Drosophila*" *Cold Spring Harbor Symp. Quant. Biol.* **62**, 1 (1997)
28. F. van Eeden and D. St. Johnston, "The polarisation of the anterior-posterior and dorsal-ventral axes during *Drosophila* oogenesis" *Curr. Opin. Genet. Dev.* **9**, 396 (1999)
29. C. Ouzounis, G. Casari, C. Sander, J. Tamames and A. Valencia, "Computational comparisons of model genomes" *Trends Biotechnol.* **14**, 280 (1996)
30. J. Tamames, C. Ouzounis, G. Casari, C. Sander and A. Valencia, "EUCLID: automatic classification of proteins in functional classes by their database annotations" *Bioinformatics* **14**, 542 (1998)