Integration of Data in Pathogenomics: Three Layers of cellular complexity and an XML-based Framework

M. Dünßer[1,2], R. Lampidis, S. Schmidt[3], D. Seipel[2] and T. Dandekar[*,1,4]

[1]Dept. of Bioinformatics, Biocenter, [2]Dept. of Informatics, Am Hubland, D-97074 Würzburg, Germany

[3]Brigham and Women's Hospital, Division of Genetics, New Research Building, Rm 464c; 77 Ave Louis Pasteur; Boston, MA 02115

[4]European Molecular Biology Laboratory, Postfach 102209, D-69012 Heidelberg, Germany

Summary

Integration of data in pathogenomics is achieved here considering three different levels of cellular complexity: (i) genome and comparative genomics, (ii) enzyme cascades and pathway analysis, (iii) networks including metabolic network analysis.

After direct sequence annotation exploiting tools for protein domain annotation (e.g. AnDOM) and analysis of regulatory elements (e.g. the RNA analyzer tool) the analysis results from extensive comparative genomics are integrated for the first level, pathway alignment adds data for the pathway level, elementary mode analysis and metabolite databanks add to the third level of cellular complexity.

For efficient data integration of all data the XML based platform myBSMLStudio2003 is discussed and developed here. It integrates XQuery capabilities, automatic scripting updates for sequence annotation and a JESS expert system shell for functional annotation. In the context of genome annotation platforms in place (GenDB, PEDANT) these different tools and approaches presented here allow improved functional genome annotation as well as data integration in pathogenomics.

1 Introduction

Integration of data in bioinformatics has to deal with numerous challenges. Particularly striking are the different levels of biological complexity, which one has to consider to enable predictions on the complex phenotype starting from the genomic sequence. We focus here on a compact prokaryotic genome from an intracellular parasite (*Mycoplasma pneumoniae*) as an application example for pathogenomics.

A bundle of tools strive to achieve integration of different results on the following three different bioinformatical levels on the road from genotype to phenotype:

(i) genome and comparative genomics

(ii) enzyme cascades and pathway analysis

(iii) networks with metabolic network analysis

* Corresponding author, phone: 0931-888-4551. FAX -4552;

e-mail: dandekar@biozentrum.uni-wuerzburg.de

We will first examine a couple of these approaches using examples from our own work. Moreover, an XML-based framework provides for us the necessary pre-requirement for optimal data integration and offers a first basis for genome-phenotype analysis allowing

- XML based data integration
- the BSML language as an adequate medium for genomic data integration
- the rule-based expert system shell JESS.
- Active re-annotation and data correction

2 Materials and Methods

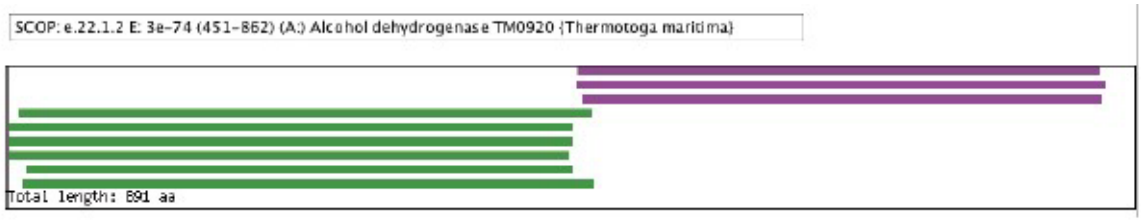
Sequences were extensively compared to available completely sequenced genomes to better assign and identify the encoded proteins therein. Furthermore, iterative sequence analysis searches compared sequences to other organisms and public databases (reviewed in Bork, 2000). The statistical expectancy value for reporting hits by chance was generally set at a conservative threshold of an expected value E of 10^{-6}. Phylogenetic analysis was applied to analyze gene duplication events and clarify the substrate specificity of the encoded enzymes. Elementary mode analysis was according to Dandekar et al. (2002). Identification and comparison of sequences to different Clusters of Orthologous Genes (COG) followed Tatusov et al. (2003). BLAST algorithm was used as described by Altschul et al. (1997). Data representation was achieved in XML. We used scripting languages (PERL) for data collection. MyBSML Studio 2003 was written in JAVA specifically for the effort (M.D.) and integrated JESS for the decision tree as well as for querying the data with KWEELT.

3 Results and Discussion

3.1 Direct sequence annotation tools for functional genomics

With the advent of large-scale sequencing techniques, many sequenced genomes are available now. This allows for comparative genomics on an extensive scale but requires also data representation of individual genomes in a suitable format.

Analyzing novel sequences from a larger sequencing effort or a complete genome involves a number of different tasks: This involves identification of transcripts (including splicing in eukaryotic genomes) as well as determination of reading frames and annotation of regulatory elements and protein domains. Helpful tools we developed to this end are the RNA analyzer to identify regulatory elements in nucleotide sequences which are RNA encoding and the AnDOM ("annotation of domains") server for the identification of structural domains in identified Open Reading Frames to examine parts of the sequence which are homologous to a known three dimensional structure (Fig. 1).




```
                                                                    Score    E
Sequences producing significant alignments:                         (bits) Value
PDB      SCOP
d1j5ra_  e.22.1.2 (A:) Alcohol dehydrogenase TM0920 {Thermotoga m...   272   3e-74
d1kq3a_  e.22.1.2 (A:) Glycerol dehydrogenase {Thermotoga maritim...   228   8e-61
d1dqsa_  e.22.1.1 (A:) Dehydroquinate synthase, DHQS {Aspergillus...   214   2e-56
d1ad3a_  c.82.1.1 (A:) Aldehyde reductase (dehydrogenase), ALDH {...   185   5e-48
d1euha_  c.82.1.1 (A:) Aldehyde reductase (dehydrogenase), ALDH {...   170   1e-43
d1a4sa_  c.82.1.1 (A:) Aldehyde reductase (dehydrogenase), ALDH {...   161   7e-41
d1ez0a_  c.82.1.1 (A:) Aldehyde reductase (dehydrogenase), ALDH {...   158   9e-40
d1bxsa_  c.82.1.1 (A:) Aldehyde reductase (dehydrogenase), ALDH {...   151   8e-38
d1ky8a_  c.82.1.1 (A:) Non-phosphorylating glyceraldehyde-3-phosp...   144   2e-35

> d1j5ra_ e.22.1.2 (A:) Alcohol dehydrogenase TM0920 {Thermotoga maritima}
          Length = 359

 Score =  272 bits (698), Expect = 3e-74
 Identities = 78/412 (18%), Positives = 150/412 (36%), Gaps = 55/412 (13%)

Query: 451 MLWHKLPKSIYFRRGSLPIALDEVITDGHKRALIVTDRFLFNNGYADQITSVLKAAGVET 510
           M    +P  ++F    L    + +   G +  ++         NG  D +  +L    +
Sbjct: 1   MWEFYMPTDVFFGEKILEKRGNIIDLLGKRALVVTGKSSSKKNGSLDDLKKLLDETEISY 60

Query: 511 EVFFEVEADPTLSIVRKGAELANSFKPDVIIALGGGSPMDAAKIMWVMYEHPETHFEELA 570
           E+F EVE +P+   V K  E   +   D ++ LGGGSPMD AK + V+ +  +   E+L
Sbjct: 61  EIFDEVEENPSFDNVMKAVERYRNDSFDFVVGLGGGSPMDFAKAVAVLLKEKDLSVEDL- 119

Query: 571 LRFMDIRKRIYKFPKMGVKAKMIAXXXXXXXXXXXXPFAVVTDDATGQKYPLADYALTPD 630
                 Y   K+     ++               P++++TD    ++        + P
Sbjct: 120 ----------YDREKVKHWLPVVEIPTTAGTGSEVTPYSILTDPEGNKRGC---TLMFPV 166

Query: 631 MAIVDANLVMDMPKSLCAFGGLDAVTHAMEAYVSVLASEFSDGQALQALKLLKEYLPASY 690
            A +D       M   L    G+DA++HA+E Y+S  ++  SD  A++A+K++   LP +
Sbjct: 167 YAFLDPRYTYSMSDELTLSTGVDALSHAVEGYLSRKSTPPSDALAIEAMKIIHRNLPKAI 226
```

Fig. 1. The structural domain server AnDOM uses PSSMs (see text) to rapidly annotate domains of known three dimensional structure in a given sequence. Result of a search for structural domains contained in the acetaldehyde dehydrogenase sequence with the AnDom Server. The colour coding reflects the predicted three dimensional structure of the domains: Green is a mixture of α-Helix plus β-Strand (SCOP Class 3, second hit, aldehyde reductase domain), blue is α/β (SCOP Class 4), violet corresponds to multi domain (SCOP class 5, top hit, alcohol dehydrogenase domain). The positions of the main structural domains are shown graphically on top, the middle displays the similarities to all protein structural domains found including the names of the domains, significance and references to sequence comparisons. Finally (bottom) the detailed alignments are given by the AnDOM server (as an example, only the start of the top alignment is shown in this figure).

Both servers are intended for smaller and larger genome annotation projects. They allow quick analysis of individual sequences identified and can also directly be installed locally to connect databases of different sequences for large scale analysis. The RNA analyzer (Bengert and Dandekar, 2003) identifies individual regulatory elements in RNA sequences using a decision tree and individual subprograms executing sequence and secondary structure searches for different elements. It exploits fast folding routines from the Vienna package (Hofacker et al., 1994). The AnDOM server (Schmidt et al., 2002) utilizes position specific scoring matrices (PSSMs) made from a large alignment of homologous sequences to individual structural domains of known experimental structure (using PDB as a reference database). Comparison of a query sequence to the PSSMs stored allows for the rapid identification of any structural domains homologous to a known structure domain according

to SCOP (LoConte et al., 2000) are in the query sequence of question. These and other tools allow rapid genome annotation and rapid identification of pathogen specific features such as host interaction factors and toxins, the results of the genome annotation are next integrated and can be made rapidly accessible using XML, BSML and specific software (see below).

3.2 Suitable bioinformatical analysis tools for three levels of biological complexity and integration

Data integration should operate on three levels of cellular complexity (Table 1):

Comparative Genomics
• annotation tools such as AnDOM (protein domains) and RNA analyzer (regulatory elements)
• identification of orthologous genes and functions encoded
• differential genome analysis and identification of species specific features
pathogenomics result: Functional inventory and data on new functional interactions
Pathway analysis
• pathway alignment
• elementary mode analysis
• calculation and identification of critical enzymes and side-routes
pathogenomics result: surprising plasticity in central pathways
Metabolic Networks
• Hub metabolites provide a helping hand
• Variable enzyme superfamilies
pathogenomics result: wide spread recruitment allows for new pathways

Table 1. The different levels of integration and specific results from our analyses

(i) *Comparative genomics*: Based on such annotated genome sequences, strict genome-to-genome comparisons allow next to identify different sets of species specific, species shared and general features for the compared genomes, in particular, pathogens and non-pathogens. This is well established and leads directly to more complex levels of data integration, notably the prediction of protein interactions by conserved genome operons (Dandekar et al., 1998), fusion events (Marcotte et al., 1999) or phylogenetic distribution of encoding genes, using tools such as STRING (von Mering et al., 2003). In *Mycoplasma pneumoniae*, a number of previously identified reading frames with unknown function could be annotated in this way by these and other comparative genome analysis techniques. Thus we could identify all proteins in the genome required for an *E. coli* like secretion pathway as an important pathogenicity factor for this intracellular parasite (Dandekar et al., 2000).

Besides functional annotation by data integration and extensive genome annotation, suitable visualization is helpful to rapidly compare and examine different genomes. To this end, dotplots of orthologous genes present a compact level of data integration to compare orientation and position of genes with similar function in different genomes (Dandekar et al., 2002).

(ii) *Pathway analysis*: On the next level, the pathway perspective, individual enzyme activities encoding protein reading frames are assembled and combined to different pathways. A concise and integrated way to compare pathway results from different genomes is

comparative pathway alignment. Species specific differences in the presence or absence of individual enzymes are identified by insertions or gaps in this comparative table (Dandekar and Sauerborn, 2002).

(iii) *Network analysis*: Finally, the different pathways are parts of cellular networks. Here the challenge is to define in a clear and mathematical way individual pathways in such a network. This is both necessary for a concise description of the network capabilities as well for the prediction of the effect of enzyme inhibitions: In many instances alternative routes through the network nevertheless allow to produce most of the metabolites. An enzyme inhibitor which can be compensated by the metabolic network is in most cases compatible with life in prokaryotic cells. Genes which encode such non-essential enzymes are themselves non-essential and lead to a non-lethal phenotype. The non-essential enzymes for a given metabolic network can easily be identified by calculation of the elementary modes. These are non-decomposable ("elementary") sets of enzymes. Each set can sustain a steady state for all internal metabolites this set of enzyme uses as substrates and products. The external metabolites used by each enzyme set need not fulfil this equilibrium condition. Using this mathematical requirement the algorithm METATOOL (Pfeiffer et al., 1999) calculates all elementary modes for a given metabolic network. Testing conditions are included so that the stable flux modes calculated are elementary, i.e., non-decomposable in sub-sets which fulfil the steady state condition for internal metabolites. This helps to answer the above questions: Any observed network flux state is always a linear combination of the elementary modes and the inhibition of a given enzyme will inhibit exactly those elementary modes in which this given enzyme occurs (Schuster et al., 2000). Recent research from our laboratory shows that this perspective of enzymes directing metabolite flows can also be turned around, metabolites shape also the way in which pathways evolve. This is strikingly shown by the observation that metabolite networks tend to be driven in structure and enzyme substrate specificities by the most frequently represented metabolites of this network. Furthermore, this helps and supports the observed wide-spread recruitment of enzymes to new pathways, allowing pathogens to rapidly change and adapt to hostile (antibiotics etc.) environments (Schmidt et al., 2003).

3.3 Automatic tool and annotation platform integration by an XML-based platform

To allow more ease in genome comparisons and achieving data integration in functional genomics one has to take into account the three different levels of complexity discussed above. We have constructed an XML-based framework as a necessary pre-requirement for optimal data integration in such genome-phenotype analysis approaches:

- XML based data integration
- BSML language as an adequate medium for genomic data integration
- the rule-based expert system shell JESS
- Active re-annotation and data correction

3.3.1 BSML as a basis for a genome annotation system

For optimal sequence analysis it is necessary to represent the incoming data in an appropriate format. These data are in most cases text (in contrast to e.g. binary data in pictures). This makes an application of XML (http://www.w3.org/XML/) advantageous. Using XML, complex and very complex text informations are easily structured and semantically represented. Current work from our laboratory focuses on "MyBSML Studio 2003". Our system is used for the automatic annotation of genes or full genomes. All data are represented

using BSML (*Bioinformatic Sequence Markup Language* (http://www.bsml.org/)). BSML is both suited for representation of sequence data (including physico-chemical properties as well as the sequence annotation. The latter are termed "Features" in BSML and represent the results of the applied expert reasoning, or, using automatic annotation, the results of the annotation tools.

Here, both results and tools for direct sequence annotation such as AnDOM or RNA analyzer (see above) as well for the integrated analysis of the next two levels of complexity such as modelling results from elementary mode analysis (e.g. participation of the annotated enzyme in the following number of elementary modes) are efficiently integrated.

The structures of the elements which contain the different results are very similar, and they can be easily processed further. BSML contains both language and information for the description and visualization of the different elements including the parameters of the applied analysis tools. Fig. 2 illustrates the correlation of XML and BSML.

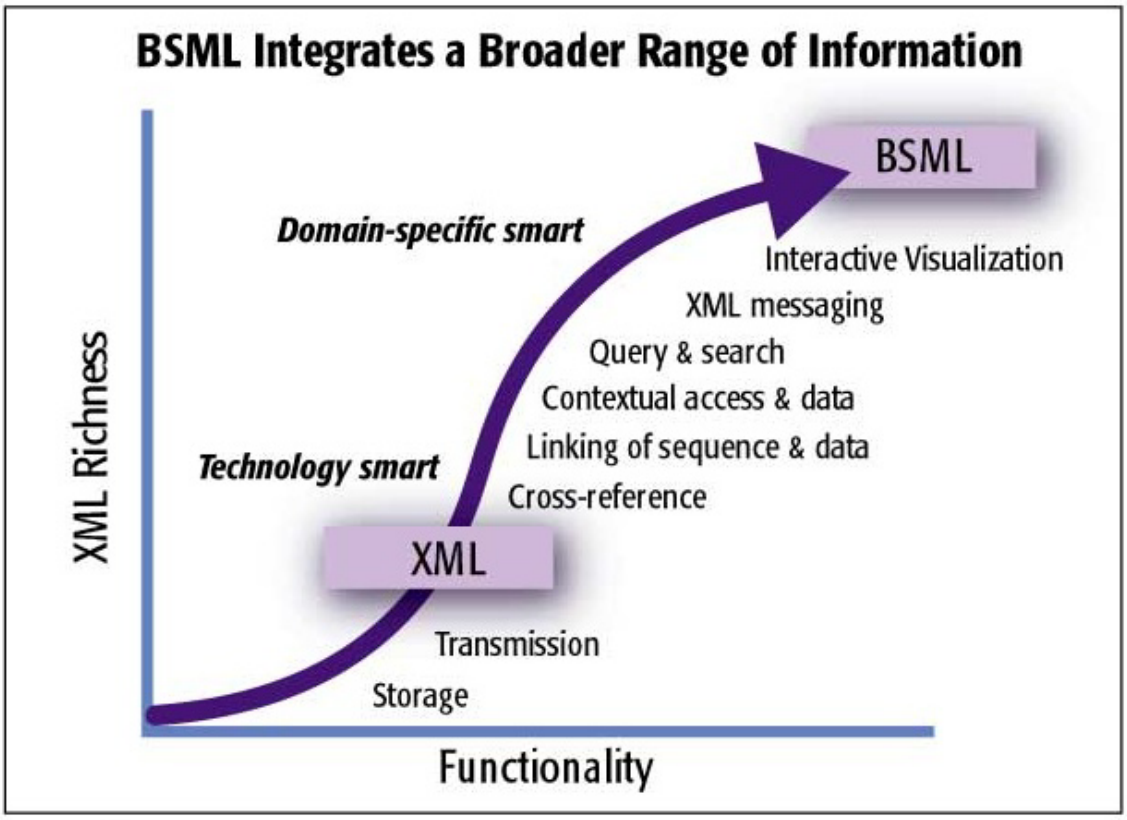


Fig. 2: BSML uses the advantages of XML and adds domain specific functionality and representation of sequence data.

BSML-documents contain data for one or several sequences and are strictly separated by first-level-elements, the "Sequence"-elements (see Fig. 3).

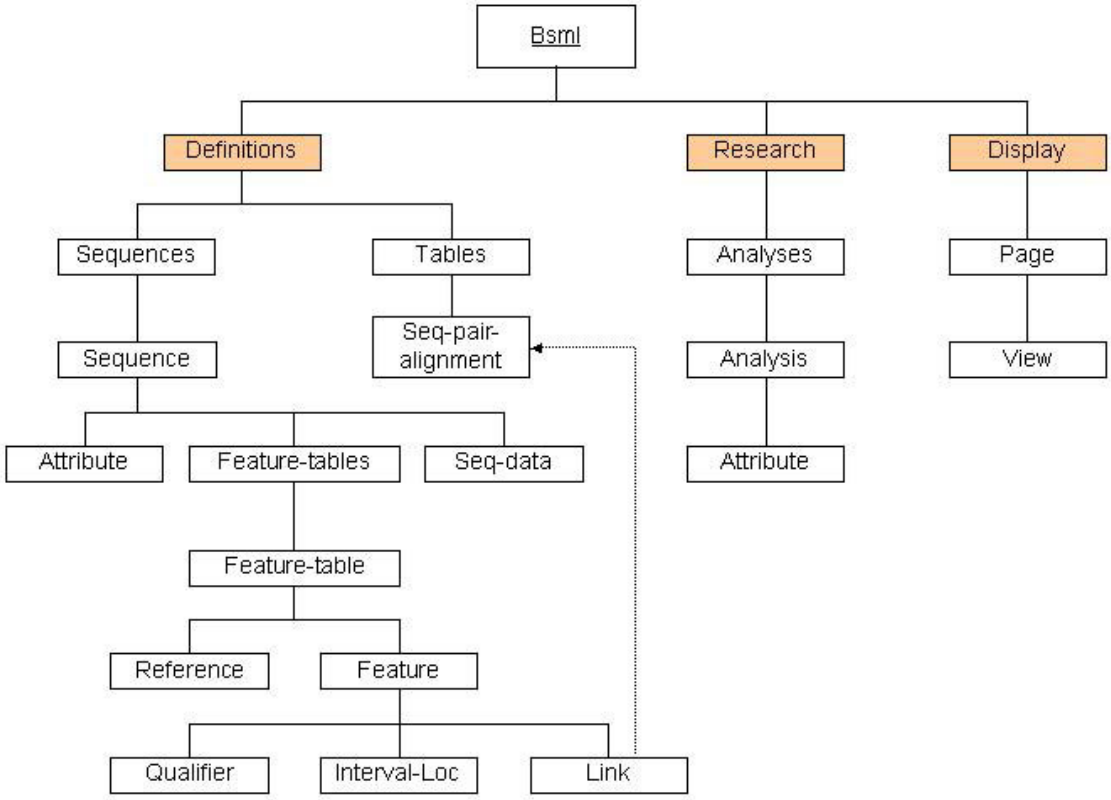


Fig. 3. Basic structure of BSML-documents. Shown are the most important elements which are also used in MyBSML Studio 2003. All Data for a sequence are stored in a "Sequence"-element which again is split into three equal parts: "Definition", "Research" und "Display".

The nature of BSML as an XML-based language enables in principle the transfer of data from any XML-supporting application and its visualization. Updates on BSML-documents are possible applying DOM, parsing is done using SAX. We are currently developing a more powerful updating tool based on the language XUpdate. Any XML-query language can be used for BSML. MyBSML Studio 2003 uses QUILT. QUILT is one of the first languages to support all language properties of XQuery 1.0 according to the standards of the World Wide Web Consortium. XQUILT is imbedded into the framework KWEELT, it supports XPath and XQuery allowing simple and complex queries (Fig. 4). In particular, this allows logical set operations on the complete genome annotation and contained proteins such as identifying all genes of *Mycoplasma pneumoniae* which are kinases and implicated as a pathogenicity factor.

```
for $a in document("MP-Table106.xml")//ENTRY
where contains($a/Comments, "kinase")
  AND contains($a/Comments, "pathogenicity factor")
return
   <RESULT>
      $a/MPN,
      $a/Gene,
      $a/Comments
   </RESULT>
```

Fig. 4: Example of a QUILT-query in the context of our annotation system:

3.3.2 Evaluation and Integration of Genome Information

Besides the easy query and access to genome data the integration and information abstraction from genome data is important. Expert systems are a powerful approach for this. We are developing an expert system module for MyBSML Studio 2003. It allows for a functional classification of an encoded protein. To achieve this, expert knowledge has to be represented as a collection of decision rules. The number of rules required depends on the classification challenge. We modelled the decision process for genome annotation as a decision tree (Fig. 5). The simplified decision tree allows already sub-classification of enzyme activities according to specific regulatory features and will assist to establish important regulatory cascades in pathogens.

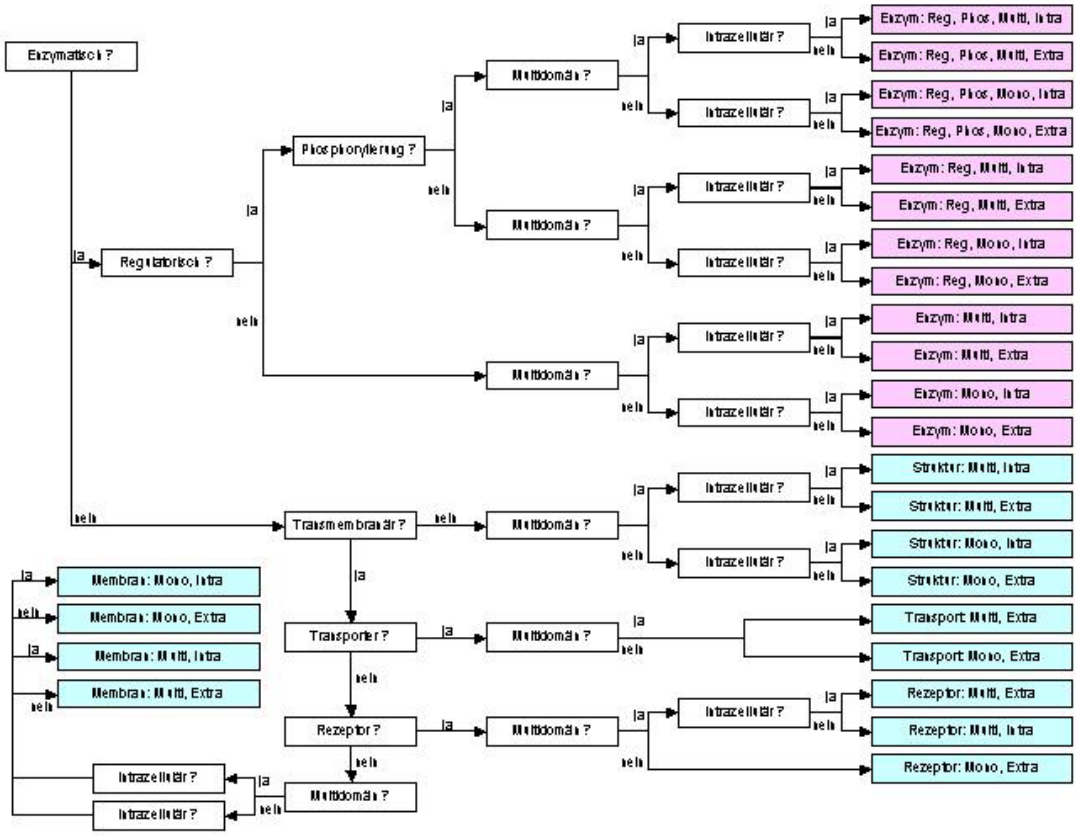


Fig. 5. Decision tree for functional classification of proteins.

The binary decision tree allows only for a simplified function classification but it can easily be extended. Each node is modelled using the same building block (Fig. 6).

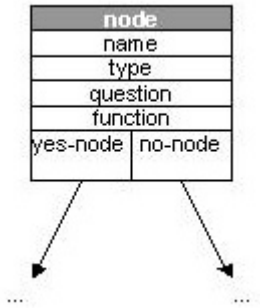


Fig. 6. Building block element for the decision tree used for function annotation in MyBSML Studio 2003.

Each node is uniquely labeled by its name and either of type "decision" (vertex node) or "answer" (end node / leaf). Only in the first case the property "function" is a pointer to an automatic annotation function to automatically assemble information for a decision. Depending on the result, different paths are followed in the decision tree until a functional classification is derived.

A test version of MyBSML Studio 2003 has already been applied to better examine the genome annotation of *Mycoplasma pneumoniae*. It contains 730 genes (688 proteins, the remainder are RNA genes). A new functional annotation combining MyBSML Studio 2003 and its automated access to automatic annotation tools is for instance the function for the genes MPN255 and MPN673: Here, new information on the protein function of this genes has recently become available from the COG database, allowing now with the system to automatically update the old function description for both proteins from "conserved hypothetical" to "2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase" (MPN673) and "4-diphosphocytidyl-2-methyl-D-erithritol-synthase" (MPN255). These enzymes participate in lipid metabolism. This shows that for these and other proteins a continuous update of their function is possible, comparing and integrating knowledge from the several annotation tools. In the concrete example the lipid metabolism of *Mycoplasma pneumoniae* is represented in a more complete fashion.

At present functional annotation with BLAST and SMART is compared to the results from COG database. Automatic collection of the results is achieved by suitable queries and automatic scripts (Fig. 5) the results are still submitted to expert evaluation. Annotation tools mentioned in the first section such as AnDOM and RNA analyzer can be directly integrated in this platform. Functional annotation itself is automated in a simple way using the java expert system shell (JESS) but will clearly need further improvements to become powerful.

After the current test phase, the annotation system will be applied to other genomes focussing on further procaryotic genomes as our department is linked to the German pathogenomics network to study genomes of bacterial pathogens.

4 Outlook

The present study shows that full data integration for functional genomics is increasingly necessary but challenging and slowly becoming feasible. We identified the different levels of genome complexity involved in this, outlined suitable software tools available and necessary for each integration step and finally discussed and tested an XML-based approach for an automatic integration of genome annotation tools and functional annotation decision making.

Our present focus is to achieve an easy comparison of different direct and higher annotation tools for specific genomes and sequences in pathogenomics. There are already nice and powerful genome annotation platforms available such as non-commercial GenDB (Meyer et al., 2003), MAGPIE (Gaasterland et al., 1996) or the commercial ones from PEDANT (Frishman et al., 2003) and BioScout (Lion Biosciences AG, Heidelberg). This should also be seen in the context of similar ongoing activities exploring the advantages for XML in bioinformatics (Achard et al., 2001) in other laboratories: There are new XML schemes for bioinformatics (Bruhn and Burton, 2003), EMBL data (Wang et al., 2002) as well as a protein markup language (Hanisch et al., 2002). An XML broker exists for integration of microarray data (Tjandra et al., 2003) and integrated systems and java editors for biological pathways (Krishnamurthy et al. 2003; Trost et al., 2003). Strong new integrated data platforms for proteomics (Aebersold et al., 2003), XML based remote procedure calls (Riva and Kohane, 2002) and a SQL-based server for online integration of life science data (Freier et al., 2002) have recently become available.

These and further softwares and platforms, in particular GenDB and PEDANT, which are both used in the department, complement and strengthen our activities to achieve optimal data integration in pathogenomics. We think that the specific approach examined and presented here, XML-based, automatic data integration for prokaryotic pathogenic genomes including identification of specific pathogenicity features is a good starting point for the challenges of data integration in pathogenomics. It integrates three levels of cellular complexity using in-house and public annotation tools, expert knowledge and an automated, continuous effort in genome annotation to create a solid basis for the inferred higher levels of cellular functions such as pathways and metabolic networks.

5 References

S. Abiteboul, P. Bunemann and D. Suciu. Data on the Web: From relations to semistructured data and XML. San Francisco: Morgan Kaufmann Publications, 2000.

ACEDB Documentation libray http://genome.cornell.edu/acedocs/

S. F. Altschul, T. L. Madden, A. A. Schaffer, 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 Research, 25(17):3389-3402, 1997.

F. Achard, G. Vaysseix and E. Barillot. XML, bioinformatics and data integration. Bioinformatics, 17(2):115-25, 2001. Review.

P. Bengert and T. Dandekar. A software tool-box for analysis of regulatory RNA elements. Nucleic Acids Research, 31(13):3441-3445, 2003.

R. E. Bruhn and P. J. Burton. Designing XML schemas for bioinformatics. Biotechniques, 34(6):1200-1206, 2003.

J. Clark, S. DeRose. XML Path Language (XPath) Version 1.0. W3C recommendation, 1999. http://www.w3c.org/TR/xpath

D. Chamberlain, J. Clark, D. Florescu, J. Robi, J. Simeon and M. Stephanescu. XQuery 1.0: An XML Query Language. W3C working draft, 2001. http://www.w3c.org/TR/xquery

T. Dandekar, B. Snel, M. Huynen and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. Trends in Biochemical Sciences, 23(9):324-328, 1998.

T. Dandekar, M. A. Huynen, J. T. Regula, B. Ueberle, C. U. Zimmermann, M. A. Andrade, T. Doerks, L. Sanchez-Pulido, B. Snel, M. Suyama, Y. P. Yuan, R. Herrmann and P. Bork. Re-

annotating the Mycoplasma pneumoniae genome sequence: Adding value, function and reading frames. Nucleic Acids Research, 28(17):3278-3288, 2000.

T. Dandekar and R. Sauerborn. Comparative genome analysis and pathway reconstruction. Pharmacogenomics, 3(2):245-256, 2002.

A. Freier, R. Hofestädt, M. Lange, U. Scholz and A. Stephanik. BioDataServer: a SQL-based service for the online integration of life science data. In Silico Biology, 2(2):37-57, 2002.

D. Frishman, M. Mokrejs, D. Kosykh, G. Kastenmuller, G. Kolesov, I. Zubrzycki, C. Gruber, B. Geier, A. Kaps, K. Albermann, A. Volz, C. Wagner, M. Fellenberg, K Heumann and H. W. Mewes. The PEDANT genome database. Nucleic Acids Research, 31(1):207-211, 2003.

T. Gaasterland and C. W. Sensen. MAGPIE: automated genome interpretation. Trends in Genetics, 12(2):76-78, 1996.

D. Hanisch, R. Zimmer and T. Lengauer. ProML--the protein markup language for specification of protein sequences, structures and families. In Silico Biology, 2(3):313-324, 2002.

I. Hofacker, W. Fontana, P. Stadler, L. Bonhoeffer, M. Tacker and P. Schuster. Fast Folding and Comparison of RNA Secondary Structures (The Vienna RNA Package). Montshefte fuer Chemie (Chemical Monthly), 125:167-188, 1994.

W. Kazakos, A. Schmidt and P. Tomczyk. Datenbanken und XML. Berlin, Heidelberg: Springer-Verlag, 2002.

L. Krishnamurthy, J. Nadeau, G. Ozsoyoglu, M. Ozsoyoglu, G. Schaeffer, M. Tasan and W. Xu. Pathways database system: an integrated system for biological pathways. Bioinformatics, 19(8):930-937, 2003.

L. Lo Conte, B. Alley, T. J. Hubbard, S. E. Brenner, A. G. Murzin and C. Chothia. SCOP: A structural classification of proteins database. Nucleic Acids Research, 28(1):257-259, 2000.

E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. Science. 285(5428):751-753, 1999.

F. Meyer, A. Goesmann, A. C. McHardy, D. Bartels, T. Bekel, J. Clausen, J. Kalinowski, B. Linke, O. Rupp, R. Giegerich and A. Puhler. GenDB--an open source genome annotation system for prokaryote genomes. Nucleic Acids Research, 31(8):2187-2195, 2003.

T. Pfeiffer, I. Sanchez-Valdenebro, J. C. Nuño, F. Montero and S. Schuster. METATOOL: For studying metabolic networks. Bioinformatics, 15(3):251-257, 1999.

A. Riva and I. S. Kohane. Accessing genomic data through XML-based remote procedure calls. Proceedings of the 2002 AMIA Annual Symposium, 662-666, 2002.

S. Schmidt, P. Bork and T. Dandekar. A versatile structural domain server using profile weight matrices. Journal of Chemical Information and Computer Sciences, 42(2):405-407, 2002.

S. Schmidt, S. Sunyaev, P. Bork and T. Dandekar. Metabolites: a helping hand for pathway evolution?

Trends in Biochemical Sciences, 28(6):336-341, 2003.

S. Schuster, D. Fell and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nature Biotechnology, 18(3):326-332, 2000.

R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin and D. A. Natale. The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 4(1):41, 2003.

C. F. Taylor, N. W. Paton, K. L. Garwood, P. D. Kirby, D. A. Stead, Z. Yin, E. W. Deutsch, L. Selway, J. Walker, I. Riba-Garcia, S. Mohammed, M. J. Deery, J. A. Howard, T. Dunkley, R. Aebersold, D. B. Kell, K. S. Lilley, P. Roepstorff, J. R. Yates 3rd, A. Brass, A. J. Brown, P. Cash, S. J. Gaskell, S. J. Hubbard and S. G. Oliver. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. Nature Biotechnology, 21(3):247-254, 2003.

D. Tjandra, S. Wong, W. Shen, B. Pulliam, E. Yu and L. Esserman. An XML message broker framework for exchange and integration of microarray data. Bioinformatics, 19(14):1844-1845, 2003.

E. Trost, H. Hackl, M. Maurer and Z. Trajanoski. Java editor for biological pathways. Bioinformatics, 19(6):786-787, 2003.

C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork and B. Snel. STRING: a database of predicted functional associations between proteins. Nucleic Acids Research, 31(1):258-261, 2003.

L. Wang, J. J. Riethoven and A. Robinson. XEMBL: distributing EMBL data in XML format. Bioinformatics, 18(8):1147-1148, 2002.