PAZAR: AN INFORMATION MALL FOR CIS-REGULATORY SEQUENCE ANNOTATION

http://www.cisreg.ca/pazar

Elodie Portales-Casamar[1], Jonathan Lim[1], Wyeth Wasserman[1], Jay Snoddy[2,3], **Stefan Kirov**[2]

1 Centre for Molecular Medicine and Therapeutics, CFRI, University of British Columbia, Vancouver, BC, CANADA
2 Graduate School in Genome Science and Technology, University of Tennessee-Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
3 Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

ABSTRACT

Gene transcription is regulated via binding of transcription factors to cis-regulatory elements (CREs) called transcription factor binding sites (TFBSs). Identification and characterization of CREs is key to intense efforts to unravel the complex regulatory programs leading to specific patterns of gene expression. Laboratory identification of CREs is both expensive and time consuming, so accurate computational predictions are desired to target laboratory resources. However, computational biologists need to have access to experimentally verified CREs, in order to test and validate CRE predicting software. Existing CRE databases seldom identify uniquely the regulatory elements. Some such databases are quite useful with respect to binding profiles (e.g. JASPAR).

PAZAR is intended as both a public database for known existing TFBSs and other CREs and as an integrated data platform to assist in the computational analysis and prediction of CREs sites. Our intention is to create a loose, multi-center framework of several dedicated and smaller databases that communicate and synchronize with a master database warehouse, which provides the minimal set of acceptance rules and allows data from different source to be compiled into a coherent datasets. In short, a compendium of boutique data collections each managed independently. In order to allow different systems to deposit data into the same database warehouse we have developed a database structure which facilitates the flexible collection of attributes while imposing a core set of methods for information extraction. We are in process of releasing an application programming interface (API) which will isolate each submission interface from the underlying database structure and reduce the complexity of the database procedures. Additionally we are developing a standardized XML file format, which could data exchange between projects or serve as a high-throughput structure for the submission and update of data.

PAZAR could benefit experimentalists by providing a more efficient means to share regulatory sequence information, thus accelerating experimental design. For computational biologists, the shared (and open-access) resource provides a richer range of regulatory sequence reference data for the assessment of predictive algorithms.

EXISTING RESOURCES

Problem: Too many disconnected databases

commercial	
TRANSFAC	eukaryotic transcription factors and their binding profiles http://www.gene-regulation.de/
transcription factors dd	transcription factors of humans and other organisms http://www.proteinlounge.com/trans_home.asp
TRRD	Transcription Regulatory Regions Db http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/
transcription factors	
JASPAR	high-quality transcription factor binding profile db http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl
RARTF	RIKEN Arabidopsis Transcription Factor db http://rarge.gsc.riken.jp/rartf/
ooTFD	object-oriented Transcription Factors Db http://www.ifti.org/ootfd/
TFdb	RIKEN Mouse Transcription Factor Db http://genome.gsc.riken.jp/TFdb/
RiceTFDB	rice genes involved in transcriptional control http://ricetfdb.bio.uni-potsdam.de/
AGRIS	AtcisDB (Arabidopsis thaliana cis-regulatory db) and AtTFDB (Arabidopsis thaliana transcription factor db). http://arabidopsis.med.ohio-state.edu/
cis-regulatory sequences	
MPD	Mammalian Promoter Db (human, mouse and rat) http://rulai.cshl.edu/CSHLmpd2
MPromDb	Mammalian Promoter Db with experimentally supported annotations http://bioinformatics.med.ohio-state.edu/MPromDb/
OMGProm	Orthologous Mammalian Gene Promoters http://bioinformatics.med.ohio-state.edu/OMGProm/
DoOP	Orthologous clusters of promoters. http://doop.abc.hu/
EPD	Eukaryotic Promoter Db http://www.epd.isb-sib.ch/
SCPD	S. cerevisiae Promoter Db http://rulai.cshl.edu/SCPD/
CEPDB	C. elegans Promoter Db http://rulai.cshl.edu/cgi-bin/CEPDB/home.cgi
PLACE	Plant Cis-acting Regulatory DNA Elements http://www.dna.affrc.go.jp/PLACE/
Plant CARE	Cis-Acting regulatory element. http://intra.psb.ugent.be:8080/PlantCARE/
PlantProm DB	Plant Promoter Sequences http://mendel.cs.rhul.ac.uk/mendel.php?topic=plantprom
OPD	Osteo-Promoter Db (promoters of genes in the osteogenic pathway) http://www.opd.tau.ac.il/
HemoPDB	Hematopoiesis Promoter Db http://bioinformatics.med.ohio-state.edu/HemoPDB/
LSPD	The Liver Specific Gene Promoter Database http://cgsigma.cshl.org/LSPD
MTIR	Muscle-specific regulation of transcription http://www.cbil.upenn.edu/MTIR/HomePage.html
the Globin Gene Server	experimental data on the regulation of the globin gene cluster http://globin.cse.psu.edu/
Oreganno	open regulatory annotation http://oreganno.org

WHAT'S NEW?

- PAZAR is an open-source and open-access data warehouse, a public repository for cis-regulatory data.
- Data is to be obtained from heterogeneous sources and then transformed to match a unified schema.
- Thus, PAZAR can be seen as an information "mall" hosting boutiques that function independently and can keep their data private or release it publicly.
- A wrapper component is placed between the individual databases and the user, presenting the data as part of one large system.
- API created as a buffer between database users and the intrinsic complexity of the schema.

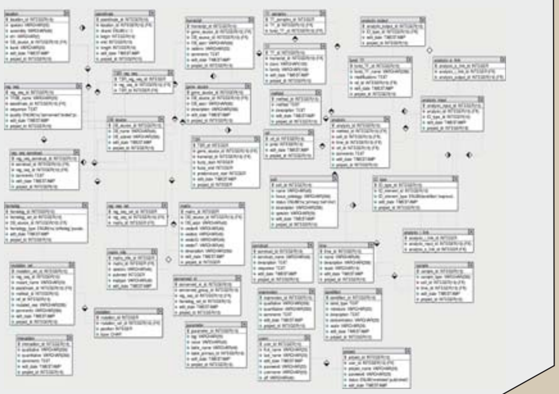
MATERIALS AND METHOD

The database model is developed through FabForce DB Designer software and is available as XML. Currently the database is implemented as a MySQL instance. The API (application programming interface) and the WI (web interface) are written in Perl and Javascript.

The API performs multiple tasks through auxiliary databases such as GeneKeyDB or EnsEMBL, which are accessed through DBD::Oracle or DBD::mysql. We use CVS as a version control system.

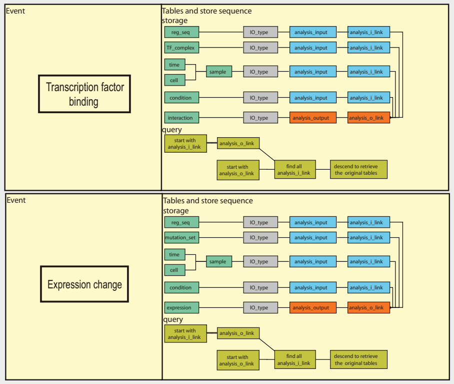
The database is currently hosted at ORNL and is protected by several backup systems (disk snapshots, mysql backup and database mirroring).

A complex database schema to allow flexibility



Storage and query of two basic events

An example of I/O link system. I/O link provides mechanism, which can store different types of relationships among events and objects (many-to-many, one-to-many and one-to-one)

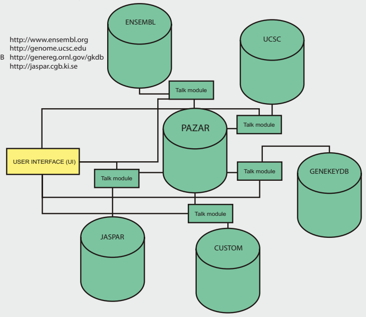


PAZAR can be linked to external data resources (ensembl, genekeydb) using a "talk" module

PAZAR is confined to the description of regulatory sequence features. There is often need for other information, such as gene identifiers, genomic DNA sequence, etc. The API talk module grants access to external resources. It is easily extensible to support other databases, including new "malls", while providing standard accessor methods.

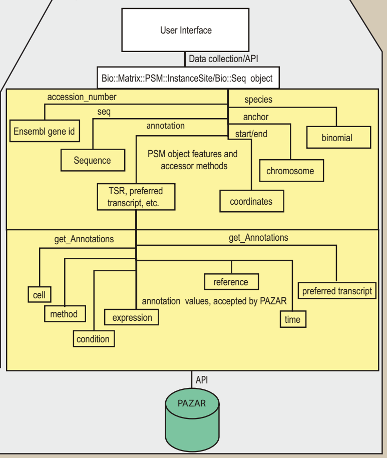
URLs:

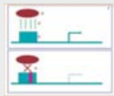
ENSEMBL	http://www.ensembl.org
UCSC	http://genome.ucsc.edu
GENEKEYDB	http://genereg.ornl.gov/gkdb
JASPAR	http://jaspar.cgb.ki.se



API data structure

The API is based on existing Bioperl data structures and methods. Using Bioperl allows the PAZAR project to use standardized procedures.





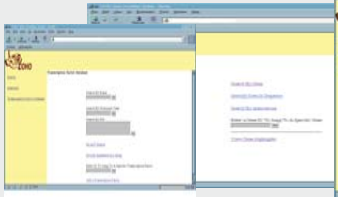
Relationship Between Entities and Tables
Entity
Tables involved

XML exchange format

```
<pazar>
  <project name="example_project" pazar_id="project_0" status="restricted">
    <user affiliation="affiliation" first_name="first_name" last_name="last_name"
pazar_id="user_0" username="username"/>
  </project>
  <data>
    <gene_source description="PDE6B" pazar_id="gene_0">
      <db_accession db_accn="ENSG00000133256" db_name="ensembl" />
      <TSR fuzzy_start="609373" fuzzy_end="609373" pazar_id="TSR_0">
        <transcript pazar_id="transcript_0">
          <db_accession db_accn="ENST00000255622" db_name="ensembl" />
        </transcript>
        <reg_seq TFBS_name="NRE" quality="tested" pazar_id="reg_seq_0"
sequence="ATTTGTAGGAGTGAGTCAGCTGACCCGC">
          <coordinate begin="609283" end="609310" length="28" strand="+">
            <location assembly="NCBI 35" band="4p16.3" species="human">
              <db_accession db_name="ensembl" />
            </location>
          </coordinate>
          <mutation_set pazar_id="mutation_set_0">
            <coordinate begin="609294" end="609299" length="6" strand="+">
              <location assembly="NCBI 35" band="4p16.3" species="human">
                <db_accession db_name="ensembl" />
              </location>
            </coordinate>
            <mutation base="g" position="1" pazar_id="mutation_0"/>
                ...
            <mutation base="a" position="6" pazar_id="mutation_5"/>
          </mutation_set>
        </reg_seq>
      </TSR>
    </gene_source>
    <gene_source description="NRL" pazar_id="gene_1">
      <db_accession db_accn="ENSG00000129535" db_name="ensembl" />
      <transcript pazar_id="transcript_1">
        <db_accession db_accn="ENST_00000250471" db_name="ensembl" />
        <tf class="bZIP" family="MAF" pazar_id="tf_0"/>
      </transcript>
    </gene_source>
    <funct_tf pazar_id="funct_tf_0" tf_ids="tf_0"/>
    <interaction qualitative="yes" pazar_id="interaction_0"/>
    <interaction qualitative="no" pazar_id="interaction_1"/>
  </data>
  <analysis >
    <method method="EMSA"/>
    <ref pmid="11438531"/>
    <cell name="Y79" species="human" status="cell_line"/>
    <input_output >
      <input inputs="funct_tf_0"/>
      <output outputs="interaction_0"/>
    </input_output>
    <input_output>
      <input inputs="mutation_set_0"/>
      <output outputs="interaction_1"/>
    </input_output>
  </analysis>
</pazar>
```

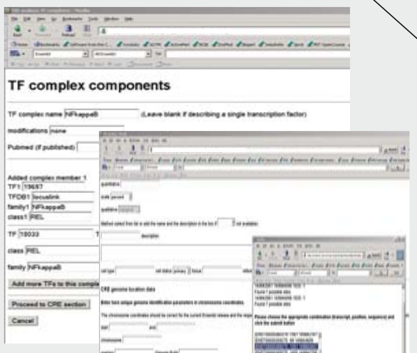
COHO gene annotation interface

An interface for collection of highly detailed information.



Streamlined web interface (TF centric)

An interface for collection of minimal regulatory sequence annotation.



TF complex components

TF complex name NFkappaB (Leave blank if describing a single transcription factor)

modifications none

Pubmed (if published)

Added complex member 1
TF1 19697
TFDB1 locuslink
family1 NFkappaB
class1 REL

TF 18033

class REL

family NFkappaB

Add more TFs to this complex

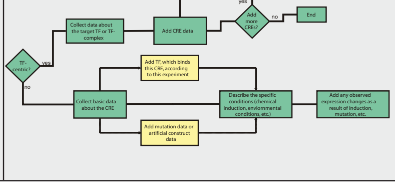
Proceed to CRE section

Cancel

CRE genomic location data

Outline of the submission algorithm

The submission process is polymorphic, which reflects the richness of experimental procedures used in CRE annotation.


TF-centric?
yes
no
Collect data about the target TF or TF-complex
Add CRE data
Add more CREs?
yes
no
End
Collect basic data about the CRE
Add TF, which binds this CRE, according to this experiment
Add mutation data or artificial construct data
Describe the specific conditions (chemical induction, enviornmental conditions, etc.)
Add any observed expression changes as a result of induction, mutation, etc.

CONCLUSIONS

- PAZAR is a public database designed to integrate gene regulatory information.
- PAZAR gives a structure to host and synchronize smaller databases each managed independently, allowing the compilation of coherent data sets from various sources.
- While the database structure is flexible enough to handle the collection of various types of data, its use will also be simplified by the continuing development of an API.
- An XML exchange format is being developed to facilitate the entry/retrieval of data to/from the database.
- The Web Interface allows PAZAR to function also as a public repository.
- Watch for updates at http://www.cisreg.ca/pazar...