Understanding the Human Genome: a Conceptual Modelling-based Approach

Prof. Oscar Pastor
ProS Research Center - Universidad Politécnica de Valencia, Spain

DEXA 2010, Bilbao
1. Why a Keynote on CM and the Human Genome?
2. Problem Statement
3. The Role of Conceptual Modeling
4. The Present
5. The Short-Term Future
6. Understanding the Domain (Problem Space)
7. Building the ER Model / Data Base (Solution Space)
8. Conclusions
We have been building
- Traditional Information Systems
- Web-based Information Systems
- SOA-based systems
- Pervasive Systems

... but, what is next?
The OO-Method Approach

Problem Space Level

- Organizational Models
- Requirements Model (Use Cases, Sequence Diagram, etc.)

Solution Space Level

- Conceptual Model
  - Object Model
  - Dynamic Model
  - Functional Model
  - Navigational Model
  - Presentation Model

- Formal Specification
  - Application Tier (COM+, EJB)
  - Persistence Tier (SQL Server, ORACLE)
  - Interface Tier (Visual Environments, Web, XML)

- Empiricism (ESE)
“A living organism is a computer or machine made up of genetic circuits in which DNA is the software that can be hacked.” — Drew Endy, MIT
Synthetic Biology can create new forms of life from scratch
- A microbe that would help in fuel production
- Biological films as a basis of new forms of lithography for assembling circuits
- Cell division counters to prevent cancer
- Re-designed seeds that the tree is programmed to grow into a house

...but, how is this “software” developed?
First synthetic cell created (announced just last month)
A tricky artificial cell
Enormously useful as a proof of concept: alive cells can be generated from genetic sequences, that could create beings with different genomes...

...provided that the genome is fully understood!!!...
Four enigmas with answer:

1. Crossing the “Rubicon” (point of no return): alive cells can be created from entirely artificial genomes
2. Bioterrorism threats
3. Does it mean creating life? Not from scratch: it is a copy of a preexistent cell
4. Will we create life? Not reason to answer no.
“Using a laptop computer, published gene sequence information and mail-order synthetic DNA, just about anyone has the potential to construct genes or entire genomes from scratch.” — Drew Endy, MIT
Model Driven Development permits

- **Reason** about the system prior to its construction
  - You can simulate the behavior to foresee the consequences of a system
- Derivate the final system in an **automatic** way
  - Obtaining a consistent result
First step: Assembling

- First abstraction step
  - Standard Biological Parts

```
#include <stdio.h>
int main(void){
  printf("hello, world\n");
  return 0;
}
```

Software
- Programming Languages
- Binary Code
  - 01010101110111
  - 01011110101011
  - 01011010010111
  - 010101111110

Life
- Standard Biological Parts
- ADN
  - gcatgctccctatcagt
gatagagattgacatc
ctatc agtgatagag
atactgacatagag
Conceptual models are needed for a systematic development of biological systems.
00010011 00000111 00000011 00001000

Physical Level

Instruction Level

Semantics: Add the values from the processor registers ‘3’ and store the result in the register ‘8’

Representation Level

3 + 4 = 7
AUG  GAA  CAC  GAC  GAG  UAA

| START | Glu | His | Asp | Glu | STOP |

Semantics: Process a protein with the four selected amino acids

However, ¿Why?
Modeling benefits are needed for biological systems

- Work at a higher abstraction level
  - Systems easy to specify
- Reason about the system prior to construction
  - Foresee consequences in advance
  - Simulate, validate, etc.
- Automate the development
  - In a systematic way
With Conceptual Models targeted at digital elements, we can improve Information Systems Development.

With Conceptual Models targeted at life we can directly improve our living.
Movement of discoveries in basic research (the Bench) to application at the clinical level (the Bedside)

A significant barrier: the lack of uniformly structured data across related biomedical domains

A potential solution: Semantic Web Technologies
Information ecosystem

- Scientific literature
- Experimental data
- Summaries of knowledge of gene products
- Diseases
- Compounds
- Informal scientific discourse and commentary in a variety of forums

This data has been provided in numerous disconnected DBs –data silos-
The lack of uniformly structured data affects many areas of biomedical research:

- Drug discovery
- Systems biology
- Individualized medicine

...all of which rely heavily on integrating and interpreting data sets produced by different experimental methods at different levels of granularity.
Still no agreement on how it is caused, or where best to intervene to treat it or prevent it
Recent hypothesis combines data from research in mouse genetics, cell biology, animal neuropsychology, protein biochemistry, neuropathology,... and other areas
Example: Huntington’s Disease (HD)

- Relatively simple genetic basis, and a model for autosomal dominant neurogenetic disorders proposed ...

- But the mechanisms by which the disorder causes pathology still not understood, what creates profound difficulties with existing treatments.
Are Semantic Web Technologies the solution?
- Thesauri, ontologies, rule systems, frame based representation systems,..
- A query language (SPARQL)
- RDF, OWL,...
Some potential advantages

- Global scope of identifiers

- RDFS and OWL are
  - Self-descriptive languages
  - Flexible, extendable and decentralized

- Ability to do inference, classification and consistency checking
  - A review of GO gave up to 10% of obsolete terms for gene annotations
Main objectives

- Identification of core vocabularies and ontologies to support effective access to knowledge and data
- Development of guidelines and best practices for unambiguously identifying resources such as docs and biological entities
- Development of strategies for linking to the information discussed in scientific pubs. from within those pubs.
Currently there are **tons of data** from the genome publicly available.

Some of these databases are **free available** on the Web because owners doesn’t know how to find relevant information.

Each database is defined with an specific schema, data format, identifications, etc.

The **integration** of the different sources is a very difficult task.
A genomic laboratory must perform an analysis to determine if the subject suffers from Neurofibromatosis.

Currently, the genetic analyst must manually search in the different databases to elaborate the report.

As a first research exercise, we have been looking for information about the NF1 Gene that provokes the Neurofibromatosis disease.

Several databases have been consulted to understand how the data is stored and retrieved.
Provides a common identification for a particular gene and the different alias used in another databases.

<table>
<thead>
<tr>
<th>Core Data</th>
<th>Database Links</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Approved Symbol</strong></td>
<td>RefSeq IDs</td>
</tr>
<tr>
<td>NF1</td>
<td></td>
</tr>
<tr>
<td><strong>Approved Name</strong></td>
<td></td>
</tr>
<tr>
<td>neurofibromin 1</td>
<td></td>
</tr>
<tr>
<td><strong>HGNC ID</strong></td>
<td></td>
</tr>
<tr>
<td>HGNC:7765</td>
<td></td>
</tr>
<tr>
<td><strong>Status</strong></td>
<td></td>
</tr>
<tr>
<td>Approved</td>
<td></td>
</tr>
<tr>
<td><strong>Chromosome</strong></td>
<td></td>
</tr>
<tr>
<td>17q11.2</td>
<td></td>
</tr>
<tr>
<td><strong>Previous Symbols</strong></td>
<td></td>
</tr>
<tr>
<td>4763</td>
<td></td>
</tr>
<tr>
<td><strong>Previous Names</strong></td>
<td></td>
</tr>
<tr>
<td>CCDS IDs</td>
<td></td>
</tr>
<tr>
<td><strong>Aliases</strong></td>
<td></td>
</tr>
<tr>
<td>CCDS11264.1</td>
<td></td>
</tr>
<tr>
<td><strong>Name Aliases</strong></td>
<td></td>
</tr>
<tr>
<td>&quot;neurofibromatosis&quot;, &quot;von Recklinghausen disease&quot;, &quot;Watson disease&quot;</td>
<td></td>
</tr>
<tr>
<td><strong>Locus Type</strong></td>
<td></td>
</tr>
<tr>
<td>gene with protein product</td>
<td></td>
</tr>
<tr>
<td><strong>Gene Symbol Links</strong></td>
<td></td>
</tr>
<tr>
<td>OTTHUMG00000132871</td>
<td>VEGA.GeneView</td>
</tr>
<tr>
<td>Ensemble ID (mapped data supplied by Ensemble)</td>
<td></td>
</tr>
<tr>
<td>ENSG00000196712</td>
<td>Ensemble.GeneView</td>
</tr>
<tr>
<td>OMIM ID (mapped data supplied by NCBI)</td>
<td></td>
</tr>
<tr>
<td>162200</td>
<td>OMIM</td>
</tr>
<tr>
<td><strong>Specialist Database Links</strong></td>
<td></td>
</tr>
<tr>
<td>UCSC ID (mapped data supplied by UCSC)</td>
<td></td>
</tr>
<tr>
<td>uc002hgg.1</td>
<td>UCSC.Index</td>
</tr>
<tr>
<td><strong>Locus Specific Database Links</strong></td>
<td></td>
</tr>
<tr>
<td>P21309</td>
<td>UniProt</td>
</tr>
</tbody>
</table>

NF1 International Mutation Database, NF1 @ The Center for Medical Genetics.
Provides a controlled vocabulary to describe gene and gene product attributes in any organism. Useful to find relationships with a particular genomic term.
Entrez Gene provides a unified query environment for genes provided by the NCBI. It can be considered as the "facto" standard database to find information about a gene.
The Human Gene Mutation Database comprises various types of mutation within the coding regions, splicing and regulatory regions of human nuclear genes causing inherited disease.
The Vertebrate Genome Annotation (VEGA) database is a central repository manual annotation of vertebrate finished genome sequence. Provides graphical views of the different gene transcripts.
A malignant peripheral nerve sheath tumor of the median arm in a patient with neurofibromatosis type 1: Report of a case.

This search has identified 12 experiments, which contain a match to your query in the title and/or protein containing a match in their name or description.

**Recent Activity**

- NFI (2332 results)
- NFI (2332 results)
- Methylglyoxal mediates p38 in human endothelia
- NFI (40668 results)
- NFI (40668 results)

**CLINICAL FEATURES**


Some patients with homozygous or compound heterozygous mutations in mismatch repair genes (see, e.g., MLH1: 129436 and MSH2: 609360) have a phenotype characterized by early onset malignancies and mild features of NFI, especially cafe-au-lait spots; see the mismatch repair cancer syndrome (G23500), sometimes referred to as brain tumor-polyposis syndrome 1 or Turcot syndrome. These patients typically do not have germline mutations in the NFI gene, although a study by Wang et al. (2005) suggested that biallelic mutations in mismatch repair genes may cause somatic mutations in the NFI gene, perhaps resulting in isolated features resembling NFI.
Navigating through hyperlinks

Tedious and repetitive

No explicit methods

Human error
Drawbacks observed

- Different identifications (ids) for the same disease gene
- The data is available on the Web but databases cannot always be directly queried
- The position (locus) of a particular gene depends on the genome sequenced
- Data is changing continuously
- High amount of information not well structured
- To provide a quality report about a gene disease several databases not interconnected must be manually consulted
The problem is getting worse !!!!!

The DNA Sequencing hardware is evolving dramatically

In next years, we will be able to sequence a complete human genome faster and cheaper
However, currently there is no software available to deal with the new challenges.

Software is required to:
- Automatically find the mutations from a sequenced sample and store the new ones detected.
- Compare the genome of different subjects in order to determine all the differences between them.
- Trace the pathway from the genome code to the final phenotype of the individuals.

Conceptual modeling is required to produce quality software in this emerging domain.
Main goal: provide Conceptual Models to represent the genome in order to enhance the Model-driven development of Biogenetic software.

The gene ontology is a useful resource to define a taxonomy but not to guide the software implementation.

The first step is to provide a common E-R model that will be able to support the genomic data complexity.

First approaches have been proposed by N.W. Paton et. al\(^1\), S. Ram\(^2\), C. Tao and D. Embley\(^3\).


The Genomic Data Chaos
The Genomic Data Chaos
Conceptual Model: Pathway View
The Input of the process is a DNA sample from a sequencing machine and an allelic reference sequence.

An alignment is performed using the BLAST tool.

Each discovered difference is formalized as an instance of the variation entity. Then, a summarized report is generated.
Authorized Variations are searched in a database conforming to the genome conceptual model.

Known variations are classified into a specific type of sequence change (Insertion, Deletion, SNP, Indel).

Unknown variations are classified as non-silent if the variation produces an effect in the expected gene product.
In order to assess the phenotype of a specific variation, a research publication is required.

The conceptual model describes the bibliographical reference that supports the phenotype of a variation.

Variations with a pathogenic phenotype are classified as mutations.

Finally, the information is gathered in a report to support the clinical diagnosis.
The entire genetic identity of an individual that does not show any outward characteristics, e.g. Genes, mutations.
Phenotype

(harder to characterise)
The observable expression of gene’s producing **notable characteristics** in an individual, *e.g.* Hair or eye colour, body mass, resistance to disease

Source: Paul Fisher - UMIST
Genotype

DNA  RNA  Protein  Protein-Protein interaction

Phenotype

Pathway  Trait

Source: Paul Fisher - UMIST
What processes to investigate?

Source: Paul Fisher - UMIST
Genes captured in microarray experiment and present in QTL (Quantitative Trait Loci) region

Phenotypic response investigated using microarray in form of expressed genes or evidence provided through QTL mapping
QTL

Gene A

Gene B

Gene C

Pathway B

Pathway linked to phenotype – high priority

Pathway not linked to phenotype – medium priority

Pathway C

Pathway not linked to QTL – low priority

Genotype
DONE MANUALLY

Pathway linked to phenotype – high priority

Pathway not linked to phenotype – medium priority

Pathway not linked to QTL – low priority

Phenotype

Genotype

Gene A

Gene B

Gene C

QTL
PubMed contains ~17,787,763 journals to date
Manually searching is tedious and frustrating
Can be hard finding the links

Computers can help with data gathering and information extraction – that’s their job!!!
Life as we know it is specified by the Genomes of the myriad organisms with which we share the planet.

The nuclear genome comprises 3,2 G nucleotides of DNA, divided into 24 linear molecules, the shortest 50M nucleotides, the longest 260M, each contained in a different chromosome.

These 24 chromosomes consist of 22 autosomes and the two sex chromosomes, X and Y

Some 35,000 genes are present in the human nuclear genome.
Understanding the Domain (the Problem Space)

**GENOME**

Transcription

**TRANSCRIPTOME**
RNA copies of the active protein-coding genes

Translation

**PROTEOME**
The cell's repertoire of proteins

Figure 1.2  *Genomes 3* (© Garland Science 2007)
Gene are made of DNA

DNA is a linear, unbranched polymer in which the monomeric subunits are four chemically distinct nucleotides than can be linked in any order and in chains containing even millions of units in length.
- Genetic code: how the nucleotide sequence of an mRNA is translated into the aminoacid sequence of a protein
- Proteins are made up from a set of 20 aminoacids
- Different sequences of amino acids result in different combinations of chemical reactivities
- Codon: codeword comprising three nucleotides
- Two-letter code is not enough, three-letter code provides 64 potential codons
- Code degeneracy
- Punctuation codons
- **Gene**: A DNA segment containing biological information and hence coding for a RNA and/or polypeptide molecule.
- **Allele**: One or two or more alternatives forms of a gene.
Building an ER Model
Genomic ER Model: Advantages

- Can be associated to different genomic databases and allows to use several gene identifications
- It has been described using terminology commonly used by biologists
- The definition of gene take into account that is not (always) a continuous sequence of bases
- The model does not include implementation details to a particular physical database schema
The Model is still to be refined and conceptually fixed...
...but it provides a solid basis to incorporate contents in a precise and structured way
... and the subsequent database can make possible an efficient use, content-oriented, where any human behaviour characteristic could be traced from fenotype to the involved gene(s)
Repairing Genetic Mutations With Lasers?

- Physical base: DNA strands differ in their light sensitivity depending on their base sequences.
- Conceptual base: need of understanding semantics behind given sequences of nucleotides

Nature versus nurture
- **Pre-implant Genetic Diagnosis**: a technique that allows to check if an embryo is/isn’t healthy from a genetic perspective, before transferred to the maternal uterus.
  - Physical base: “assisted reproduction” technologies
  - Conceptual base: need to understand semantics of specific gene mutations
Discovered a gene – **EYS** (for “Eyes Shut”) that causes *inherited blindness*.

- Physical base: mutation that gives rise to the problem
- Conceptual base: why the mutation occurs? How to prevent it?
- Identified **295 potential therapeutics targets against AIDS**
  - Physical base: 295 human proteins that “probably” helps the AIDS to establish in the human cells
  - Conceptual base: “probably”? Under which conditions / interactions?
Understanding the Human Genome can become an extremely hard task if research is more and more oriented to the solution space.

Discovering “human” patterns in the genomic code is really like looking for a needle in a haystack.

Conceptual Modeling-based approaches and techniques applied to this challenging domain should guide the efforts to succeed.
And more and more challenges to be explored...

- Linking diseases with genes with therapeutical purposes as a main application
- Gene mutations that enforce expression of some other genes while delaying or reducing the expression of others
- Gene regulators
Conclusions

Una polla xica, pica, pellarica, camatorta i becarica...

Immune system
Ontology
Cytosine
Terminator
Genes against the malaria
Allele
Centromere
Aminoacid
OWL
Hydrogene bonds
Exon skipping
Nucleotides
Vertebrate Genome Annotation
MAJOR groove

Base pair
Transcription
Exon
Human
ORF
Chromosome
Regulator sequence
DNA
RNA polymerase
Neutral
polimorphism
Regulator sequence
Diagnosis
Mutation
Chromosomic mutation
Exon skipping
Intergenic region
RDF

Protein
Gene
Spliced transcript
Transcription unit
Chromosomic mutation
Terminator
Allelic variant
Gene Ontology

Analysis
Cell
Intron
Nature versus nurture
DNA
ORF
Gene Ontology

Data bank
ORI
Genome

Proteone
External identification
Telomere

Adenine
Thymine
Guanine

HUGO
Pre-implant genetic diagnosis
Embryo
Entrez Gene

Research centre
Enhanced sequence
Ambient

Human Gene Mutation Database
Repairing genetic mutations with lasers
An 'infidelity' gene for men

'Fat' gene makes greedy

Conceptual Modeling-based
Conceptual model
OO-Method

Genetic influences on female infidelity

Human Gene Mutation Database
This is probably the most attractive challenge in the future of the Conceptual Modeling community:

*Modeling the Real Life to understand why we are as we are, and how a human being can be seen as the “representation” of a Conceptual Model that can be specified in detail*
Thanks for your attention!