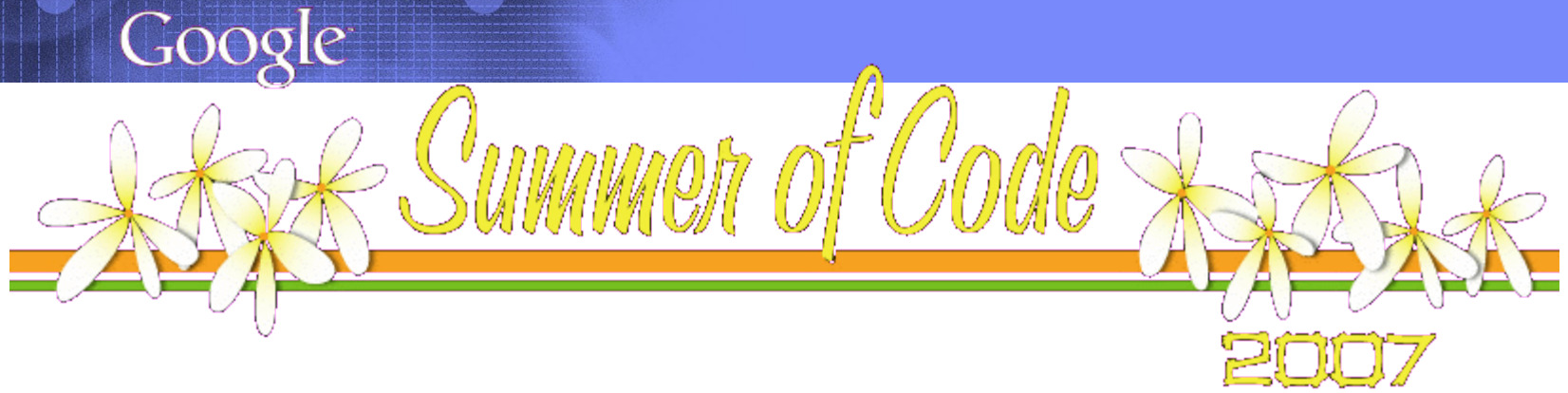


Project: GeneQuad

Intern: Daniel Lélis Baggio

Mentors: Dr. Simon Lin and Dr. Pan Du

Agenda

- Summer of Code overview
- Intro project: GenDex
- GenQuad
- Research findings and results
- Comments or questions

Summer of Code “flip bits, not burgers”

- Google sponsored Summer internship for student developers to write code for various open source projects
  - 2005 – 8,000 proposals, 419 positions, 39 organizations
  - 2006 – 6,000 proposals, 600 positions, 102 organizations
  - 2007 – 6,200 proposals, 900 positions, 131 organizations
- Goals
  - Get more open source code created and released for the benefit of all

GenDex: A study of the popularity of genes

- Goal
  - Find out the most popular genes of each year through an index based on the number of published geneRifs
- What's GeneRIF?
  - Gene Reference Into Function (255 character description)
  - NCBI indexers/ community based
- Procedures
  - Importing geneRif to a relational database (sqlite3)
  - Workarounds (multiple PubMedID for same geneRIF)
  - 162,725 records generated ~ 10 minutes @ PentiumM

GenDex: A study of the popularity of genes

- Imported GeneRIF database layout

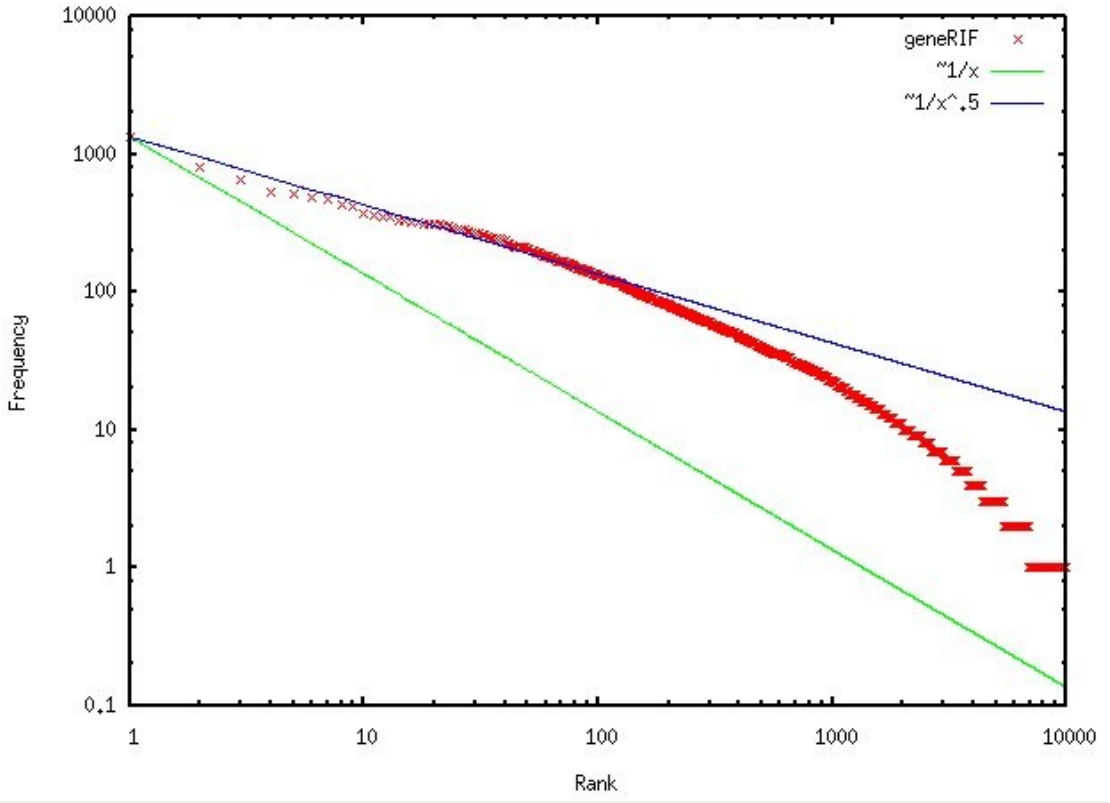
id	taxID	geneID	pmID	updateDay	updateHour	description
1	9606	708	1830244	2004-05-11	13:09	This protein has also been identified as the p32 subunit of pre-mRNA splicing factor SF2, as well as a hyaluronic acid-binding protein

- **Tax ID:** the unique identifier provided by NCBI Taxonomy for the species or strain/isolate
- **Gene ID:** the unique identifier for a gene
- **PubMed ID (PMID):** unique citation identifier(s) in PubMed;
- **GeneRIF text:** GeneRIF text string, length <= 255 characters

Zipf's Law

- Empirical law stating that:
  - “In a corpus of natural language utterances, the frequency of any word is roughly inversely proportional to its rank in the frequency table”
- Brown Corpus (1 million words):
  - “The” (7 %) ~ 1/1
  - “Of” (3.5%) ~ 1/2
  - “And” (2.7%) ~ 1/3
- Why is it important?
  - 135 words cover half of Brown Corpus
- What about GeneRIFs?

Zipf's Law



- GeneRIFs follow $\frac{1}{\sqrt{Rank}}$ proportion

Coverage Tests

Number of genes with highest frequency	Number of geneRIFs covered	Percentage of human genes covered(%)
1	1353	1.3
100	24901	23.6
500	51493	48.8
1000	66372	62.9
5000	98238	93.1

- Human genes covered by GeneRIF = **9,984**
- GeneRIFs related to human genes = **105,516**
- **50%** of studies (GeneRIFs) focused on **5%** of genes

Which are the most studied genes?

- GeneRIF does not give much description about genes
  - Importing Entrez gene_info database
  - Similar procedure, but:
    - gene_info.db ~ 300 MB
    - 12 hours @ PentiumM
  - Generated description of the 20 most studied genes in the documentation

Which are the most studied genes?

taxID	9606
geneID	7157
symbol	TP53
synonyms	LFS1 TRP53 p53
dbXrefs	HGNC:11998 MIM:191170 HPRD:01859
chromosome	17
mapLocation	17p13.1
description	tumor protein p53 (Li-Fraumeni syndrome)
typeofGene	protein-coding
symbolAuthority	TP53
fullNameAuthority	tumor protein p53 (Li-Fraumeni syndrome)
otherDesignations	p53 tumor suppressor | tumor protein p53

- Most studied gene is **cancer** related

Gathering information about publication time

- GeneRIF does not point to the exact time genes were being studied, but points to PubMED articles
- Solution: using BioRuby

- Integrated environment for Bioinformatics with Ruby language
- Ruby scripts to fetch PubMED ~ 6 hours

Rising Stars

- Finding genes which got more attention from year to year
- Classifier (C++) program to rank and calculate rising stars
- Ordered by total percentage change

2006 Rising Stars

GeneID	Percentage change(%)	Rank in 2005	Rank in 2006	GeneRIFS in 2005	GeneRIFS in 2006	Gene description
3075	+0.12	364	58	10	37	complement factor H
1029	+0.11	26	10	53	79	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
472	+0.11	97	35	26	51	ataxia telangiectasia mutated (includes complementation groups A, C and D)
3717	+0.11	104	38	25	49	Janus kinase 2 (a protein tyrosine kinase)
154	+0.08	116	50	23	42	adrenergic, beta-2-, receptor, surface
5243	+0.08	46	17	45	64	ATP-binding cassette, sub-family B (MDR/TAP), member 1

- Rising stars for 2002, 2003, 2004, 2005, 2006 and 2007 can be found in the project wiki

2006 Show Stoppers

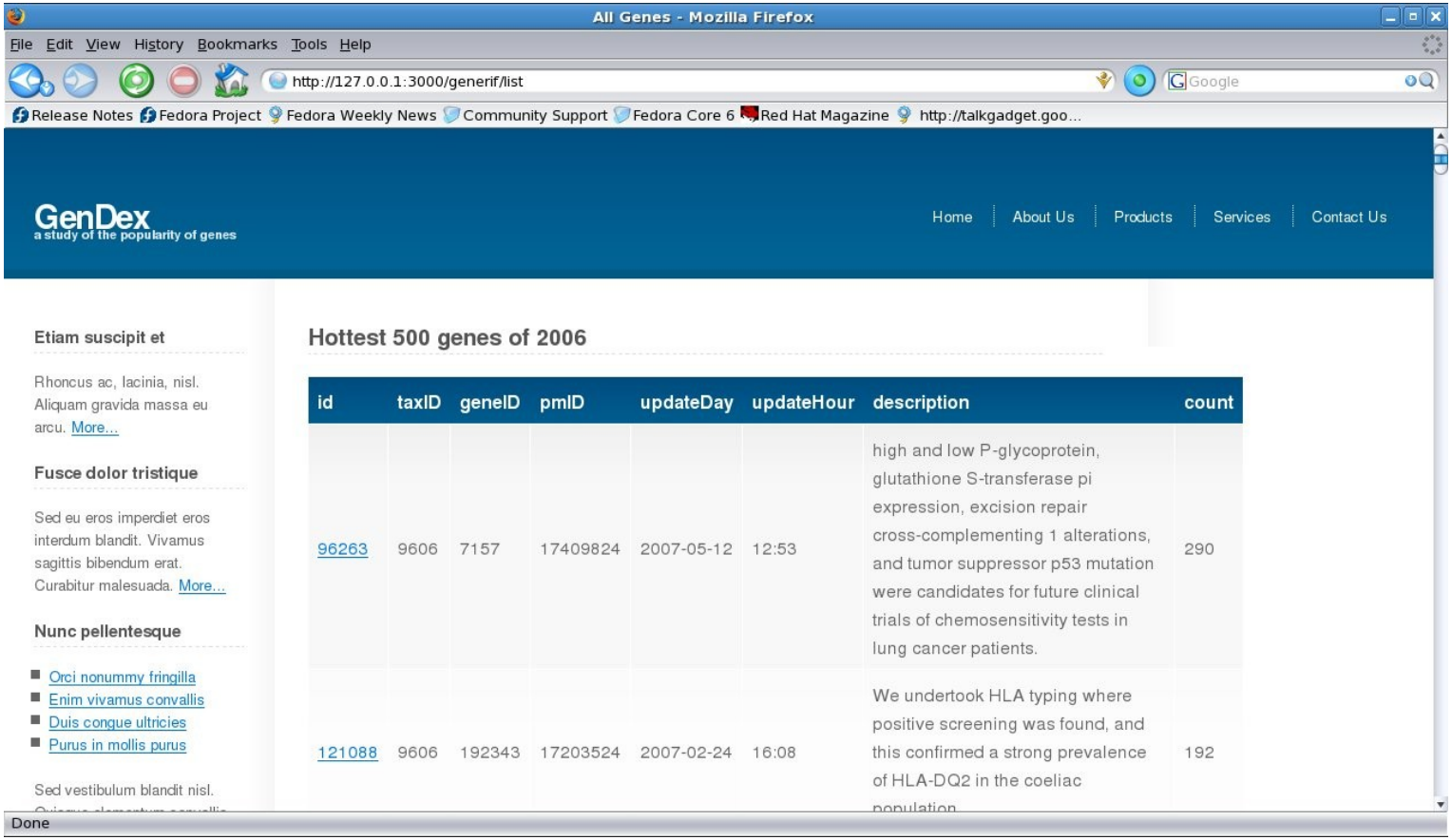
- Finding genes which were forgotten along the years

GeneID	Percentage change(%)	Rank in 2005	Rank in 2006	GeneRIFS in 2005	GeneRIFS in 2006	Gene description
1026	-0.12	13	44	69	45	cyclin-dependent kinase inhibitor 1A (p21, Cip1)
3952	-0.12	20	65	58	34	leptin (obesity homolog, mouse)
1027	-0.09	42	106	46	27	cyclin-dependent kinase inhibitor 1B (p27, Kip1)
5925	-0.09	38	90	47	29	retinoblastoma 1 (including osteosarcoma)
3576	-0.09	12	29	71	54	interleukin 8
4790	-0.09	5	8	102	86	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (p105)

- Static tables reported in project wiki

AJAX Ruby on Rails project running on NU server

- http://rails-dev.bioinformatics.northwestern.edu/app-name



GeneQuad

- Automatically finding 4 words that best describe a gene based on GeneRIF studies
- Web application with user friendly interface
- Natural Language Processing

GeneQuad - Stemming

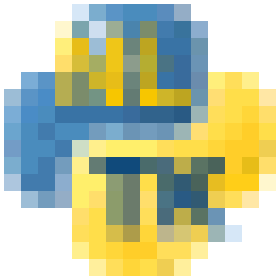
- Stemming is the process for reducing inflected words to their stem, base or root form
  - testing -> **test**
  - companies -> **compani**, company -> **compani**
- Google search adopted stemming in 2003
- Using Porter's algorithm from Python Natural Language Toolkit
- After installing:

```
>>> from nltk import stem
>>> p = stem.porter.Porter()
>>> p.stem('testing')
'test'
```

GeneQuad - Stemming

- Frequency of stemmed words for all GeneRIF descriptions

GeneRIF Stemmed word	Frequency	Brown Corpus word	Frequency
of	205615	the	69967
the	175063	of	36406
in	164756	a	34647
and	152792	and	28855
a	71672	to	26142
to	62849	in	21338
is	56088	that	10774
cell	49655	is	10093



that	43647	was	9808
with	42050	he	9794
by	39557	for	9485
express	36867	it	9047
for	36045	with	7286
activ	35625	as	7250
role	26650	his	6994
protein	26007	on	6735
gene	24169	be	6373
regul	24034	at	5377
may	19123	by	5303
receptor	17785	this	5138

GeneQuad – Filtering (tf.idf term weighting)

- Inverse document frequency (the more unique, the higher)

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d : d \ni t_i\}|}$$

- weight (term i, document j) = term frequency * idfi

- Another approach – Brown Corpus word ratios

GeneQuad – Results

- geneID = 39
- Fully differentiated macrophages express ACAT2 in addition to ACAT1 under various pathologic conditions.
- ACAT-1 transcripts predominate in **human liver** and ACAT-2 transcripts predominate in human duodenum and support the notion that ACAT-2 has an important regulatory role in liver and intestine.
- ACAT2 provided the major **cholesterol-esterifying** activity in 3 of 4 human liver samples.
- Data describe the high resolution structure of human cytosolic acetoacetyl-CoA **thiolase** (CT), both unliganded (at 2.3 angstroms resolution) and in complex with CoA (at 1.6 angstroms resolution).
- transcription factors hepatic nuclear factor 1 (HNF1)alpha and beta play an important part in the regulation of the human **acetyl-Coenzyme** A acetyltransferase 2(ACAT2) promoter
- The structural features of various sterols as substrates and/or activators of ACAT1 and ACAT2 in vitro are reported.
- Elevated ACAT2 expression may serve as a new biomarker for certain form(s) of hepatocellular carcinoma.
- Alternative splicing produces two human ACAT2 mRNA variants that encode the novel ACAT2 isoenzymes. Our findings might help to understand the regulation of the ACAT2 gene expression under certain physiological and pathological conditions.
- histidine residues located at the active site are very crucial both for the catalytic activity of the enzyme and for distinguishing ACAT1 from ACAT2 with respect to enzyme catalysis and substrate specificity

GeneQuad – Results

- Words and weights

Gene 39	
acat2	70.330
acat1	23.666
acat-2	18.134
predominate	16.513
resolution	12.871
angstroms	12.685
human	11.687
cholesterol-esterifying	10.454
liver	9.830
acetoacetyl-coa	9.760
acat-1	9.355
certain	9.220
thiolase	8.844
transcripts	8.388
acetyl-coenzyme	8.056
various	8.048
hnf1	7.745
conditions	7.424
sterols	7.409
promoterthe	7.158
under	7.023
distinguishing	7.020
isoenzymes	7.020
duodenum	6.988
unliganded	6.843
coa	6.790

Project info (google for “genequad”)

- Project homepage
  - http://code.google.com/p/genequad/
- Documentation
  - http://code.google.com/p/genequad/wiki/ActivityLog
- Subversion repository
  - http://genequad.googlecode.com/svn/
- Contact
  - danielbaggio@gmail.com

Comments / Questions