

ProMog CellGrams

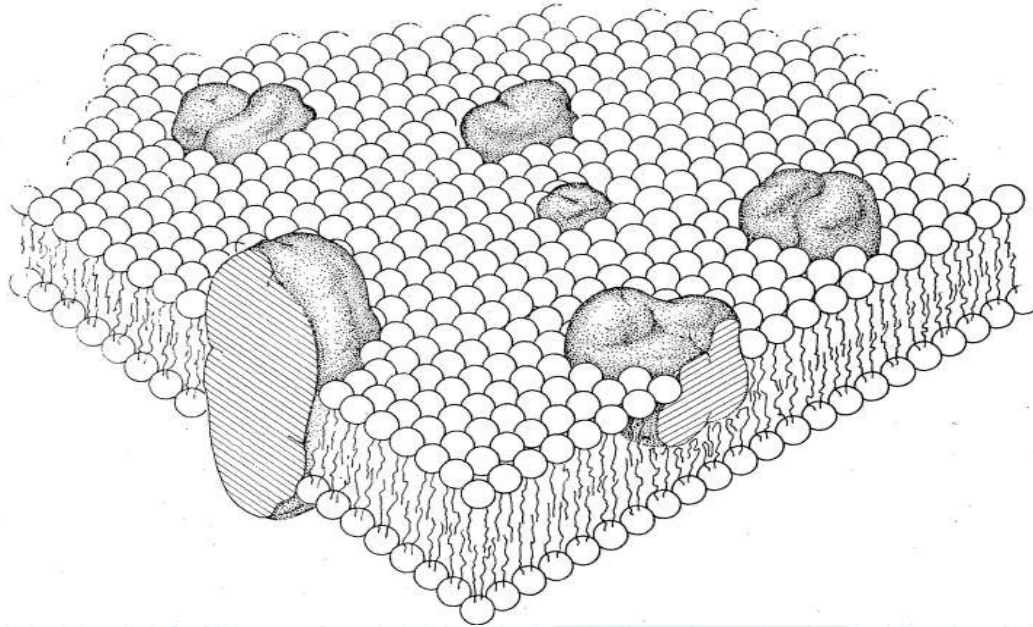
Evidence-Based Proteome-Wide
Subcellular Location – A Basis
Partitioning for Classification Systems

Kayven Riese

Overview

- Theoretical basis of subcellular location (SCL) and transmembrane proteins (TMPs)
- Support in medical research for compartmental modeling & Fluid Mosaic Model (FMM)
- TMP topology and Ahram et al SW analysis
- Universal Protein Resource (UniProt) Releases
- UniProt utilization performance studies
- ProMog.c Proteomic Demographics module
- Cellgram.c diagram module & Cairo Graphics library
- Wolfram Alpha online math service and text rendering
- Results of Study

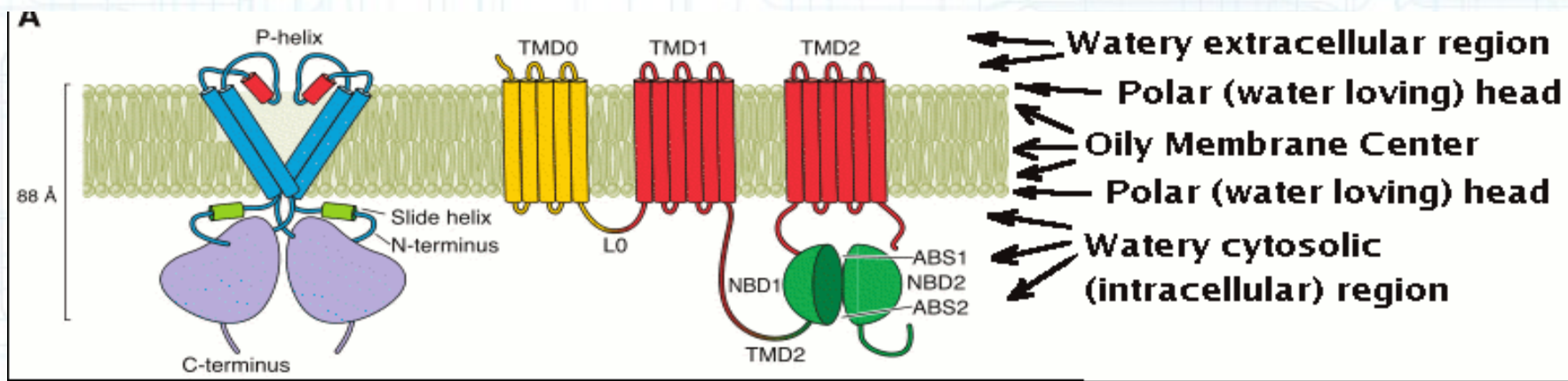
Fluid Mosaic Model (FMM)



Proposed by Singer and Nicholson in 1972

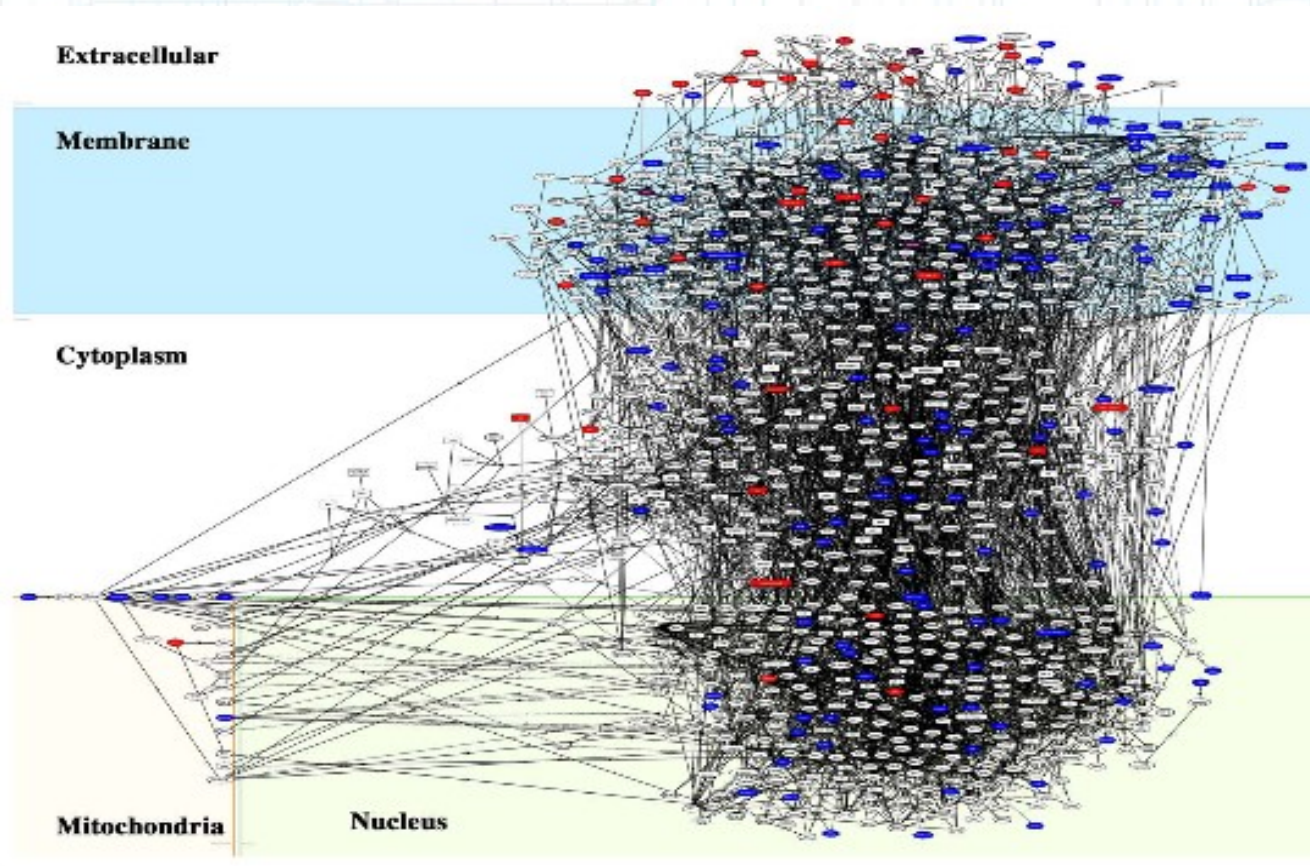
- Theory states some proteins are lodged in membranes
- Oily membranes separate watery compartments
- Special proteins constrained in psuedo-2D
- Serve as gatekeepers between cellular regions

Internal Structure of Transmembrane Proteins (TMPs)



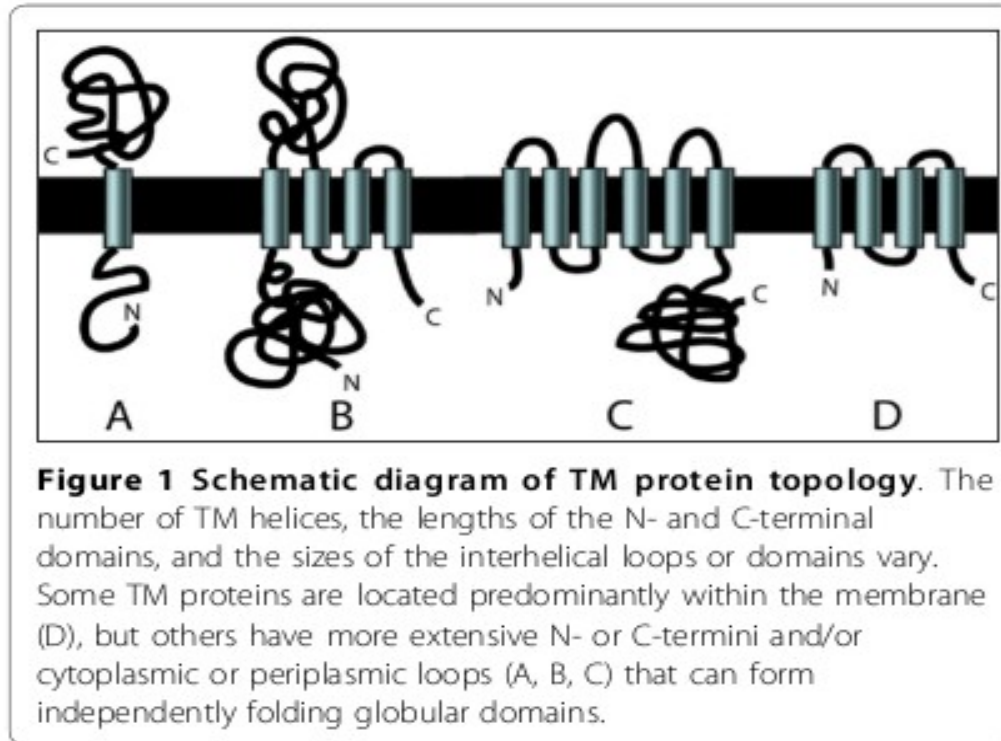
- Chemical interplay of oily and watery molecules
- “Amphiphilic” molecules have both properties
- TMPs form barrier between life and death

Medical Research Groups Have Proposed 4 & 5 Compartment Protein/Gene Networks



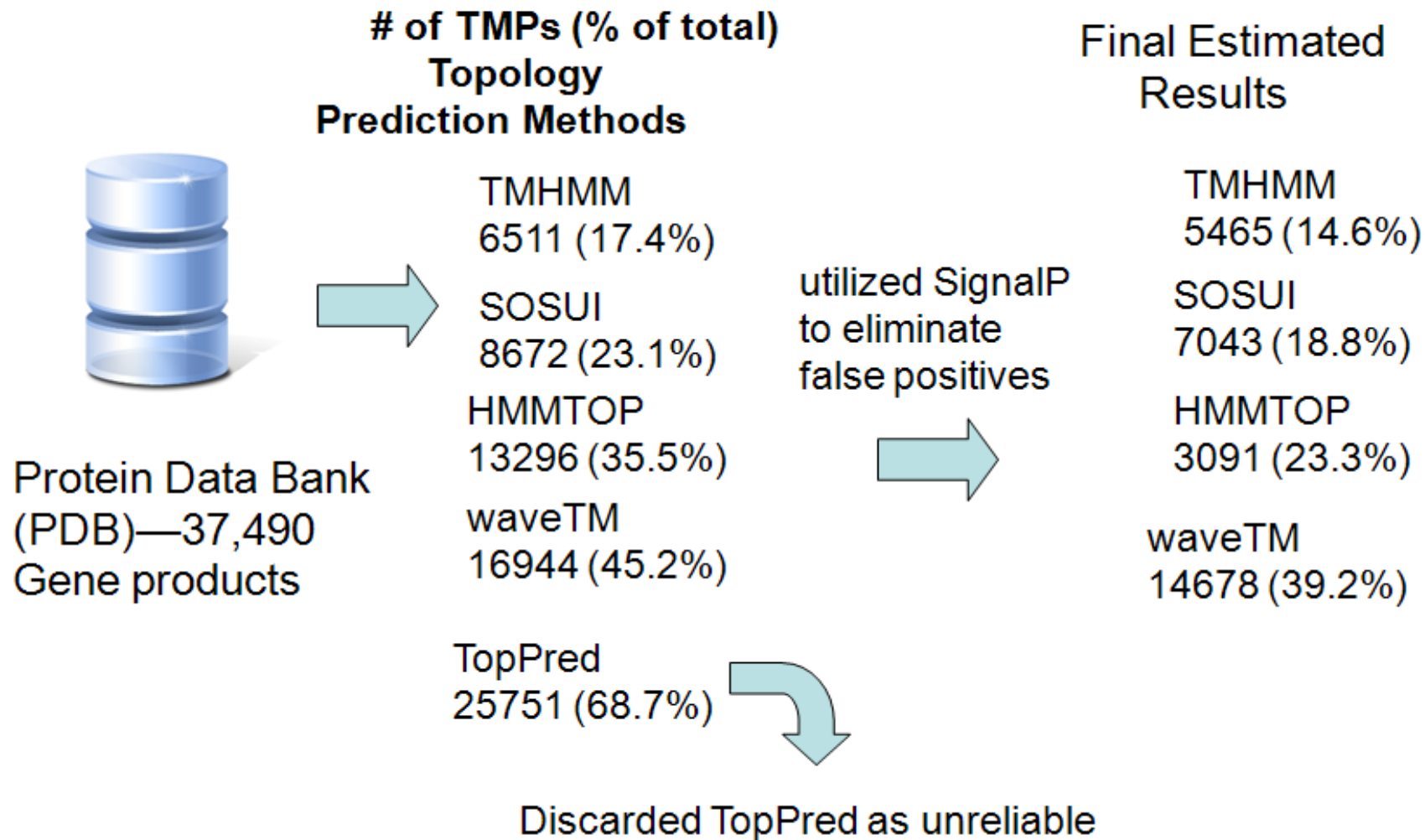
- Five Compartment model includes mitochondria
- Four Compartment model employed w/o mitochondria

TMP Topology



- Transmembrane Protein (TMP) threaded in membranes
- Some TMP sub-sequences transmembrane
- Other sub-sequences belong to various compartments

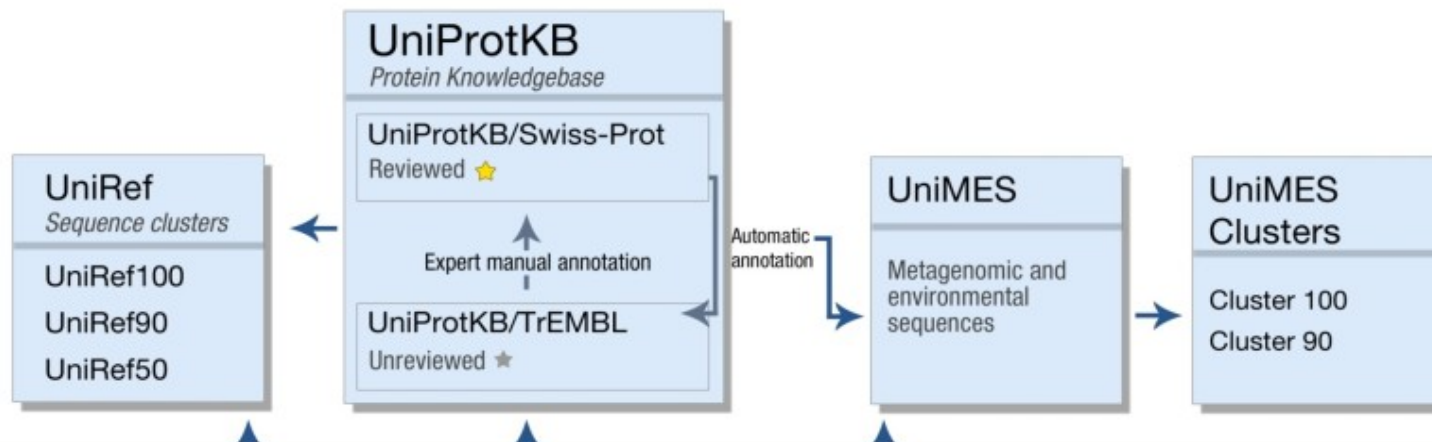
How Ahram et. al. Estimated Total TMPs in Human Genome



The Universal Protein Resource (UniProt)

About UniProt

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the [UniProt Knowledgebase \(UniProtKB\)](#), the [UniProt Reference Clusters \(UniRef\)](#), and the [UniProt Archive \(UniParc\)](#). The [UniProt Metagenomic and Environmental Sequences \(UniMES\)](#) database is a repository specifically developed for metagenomic and environmental data.



UniProtKB/Swiss-Prot is the manually annotated smaller version compared To UniProtKB/TrEMBL

UniProt Updates

The screenshot shows a web browser window with the address bar displaying <http://www.uniprot.org/news/?query=UniProtKB/Swiss-Prot&sort=score>. The browser's tab bar shows several open tabs: "ExPASy - Data...", "Download", "UniProtKB/Swis...", "Taxonomy", and "UniPr...". Below the tabs, the page content displays "106 results for UniProtKB/Swiss-Prot in News sorted by score descending". The results are listed in descending order of score, with the most recent release at the top. The visible results are:

- TrEMBL release 24.0 - June 1, 2003**
- TrEMBL release 26.0 - March 2, 2004**
- UniProt release 2011_05 - May 3, 2011**
Complete proteome sets for Homo sapiens and Mus musculus
Statistics for UniProtKB: [Swiss-Prot](#) · [TrEMBL](#)
- UniProt release 14.0 - July 22, 2008**
Major release · New official UniProt website · New structure for DE lines · Cross-reference to BindingDB · UniProt c
- UniProt release 7.0 - February 7, 2006**

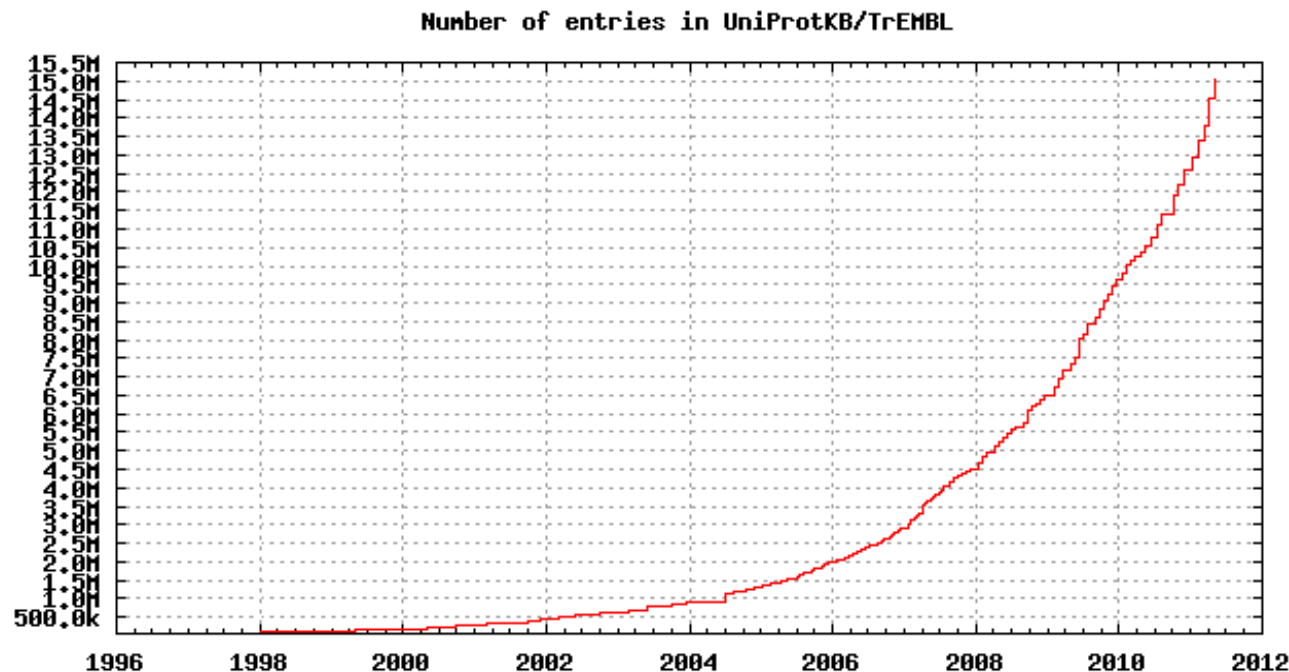
UniProt periodically updates release data

Exponential Growth of UniProt TrEMBL Data

EBI > Databases > Protein > UniProt > UniProtKB/TrEMBL

UniProtKB/TrEMBL - Current Release Statistics

The growth of the database is summarized below.



Over the past decade, data has exponentially increased

TrEMBL File Data Load

```

human_REMAINDER.txt
human_subc_loc.txt
huma.png
images
late_where.png
Makefile
med_prot.dat
minimini_prot.dat
mini_prot.dat
more_wheres.png
muscle.2011.03.24.14.54.png
muscle.2011.03.24.16.30.png
muscle.2011.03.24.17.11.png
muscle.2011.03.24.17.22.png
muscle.2011.03.24.18.02.png
[kayve@kayve-centOS integrated]$ gunzip uniprot_trembl.dat.gz
[kayve@kayve-centOS integrated]$ ./promog uniprot_trembl.dat

total.2011.03.24.20.51.png
total.2011.03.24.22.22.png
total.2011.04.21.11.48.png
total.2011.05.01.23.21.png
total.2011.05.01.23.23.png
total.2011.05.04.21.50.png
total.2011.05.04.23.41.png
total.2011.05.04.23.53.png
total.png
tota.png
uniprot_sprot.dat
uniprot_trembl.dat.gz
variables
waiter.bash
where_everything.png

uniprot_trembl.dat
root@kayve-centOS ~]# ls -l /home/kayve/thesis/integrated/uniprot_trembl.dat
-rw-r--r-- 1 kayve kayve 37526760165 May  5 03:05 /home/kayve/thesis/integrated/uniprot_trembl.dat
root@kayve-centOS ~]#

```

```

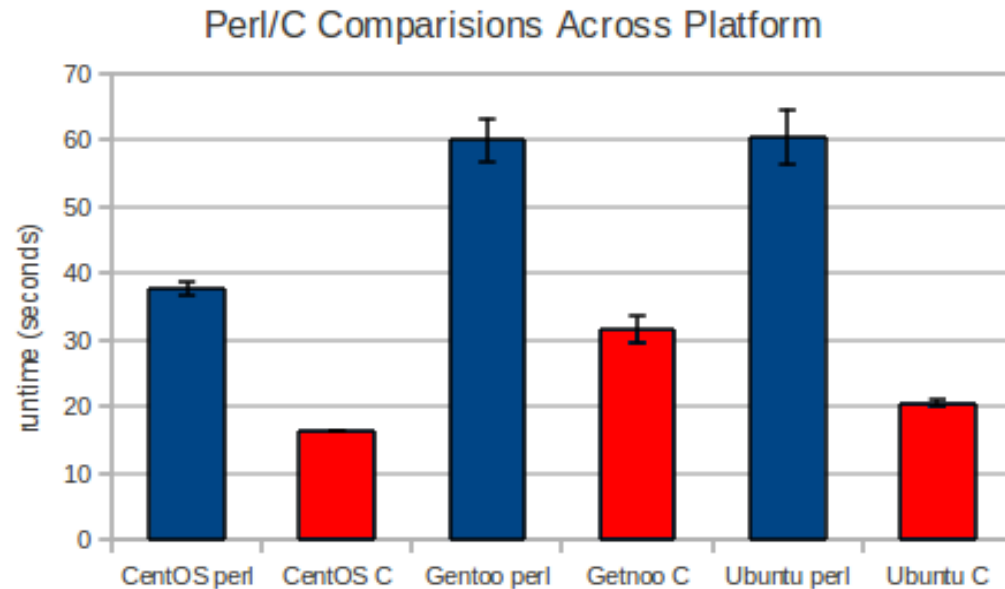
File Edit View Terminal Tabs Help
[root@kayve-centOS ~]# df -h
Filesystem              Size  Used Avail Use% Mounted on
/dev/mapper/VolGroup00-LogVol00
                        446G  369G   54G   88% /
/dev/sda1                99M   33M   62M   35% /boot
tmpfs                   1.9G     0  1.9G    0% /dev/shm
You have new mail in /var/spool/mail/root
[root@kayve-centOS ~]# du /home/kayve/thesis/integrated/
508    /home/kayve/thesis/integrated/times_scripted_jun2
30543768 /home/kayve/thesis/integrated/old.dat
16460   /home/kayve/thesis/integrated/images
66157584 /home/kayve/thesis/integrated/
[root@kayve-centOS ~]# du /home/kayve/thesis/integrated/
508    /home/kayve/thesis/integrated/times_scripted_jun2
30543768 /home/kayve/thesis/integrated/old.dat
16460   /home/kayve/thesis/integrated/images
95694480 /home/kayve/thesis/integrated/
[root@kayve-centOS ~]# df -h
Filesystem              Size  Used Avail Use% Mounted on
/dev/mapper/VolGroup00-LogVol00
                        446G  398G   25G   95% /
/dev/sda1                99M   33M   62M   35% /boot
tmpfs                   1.9G     0  1.9G    0% /dev/shm
[root@kayve-centOS ~]#

```

- Current TrEMBL file contains 37GB of Data
- Represents Significant System Load
- High performance computing called for
- Desired calculation a minute fraction of full *in silico* simulation

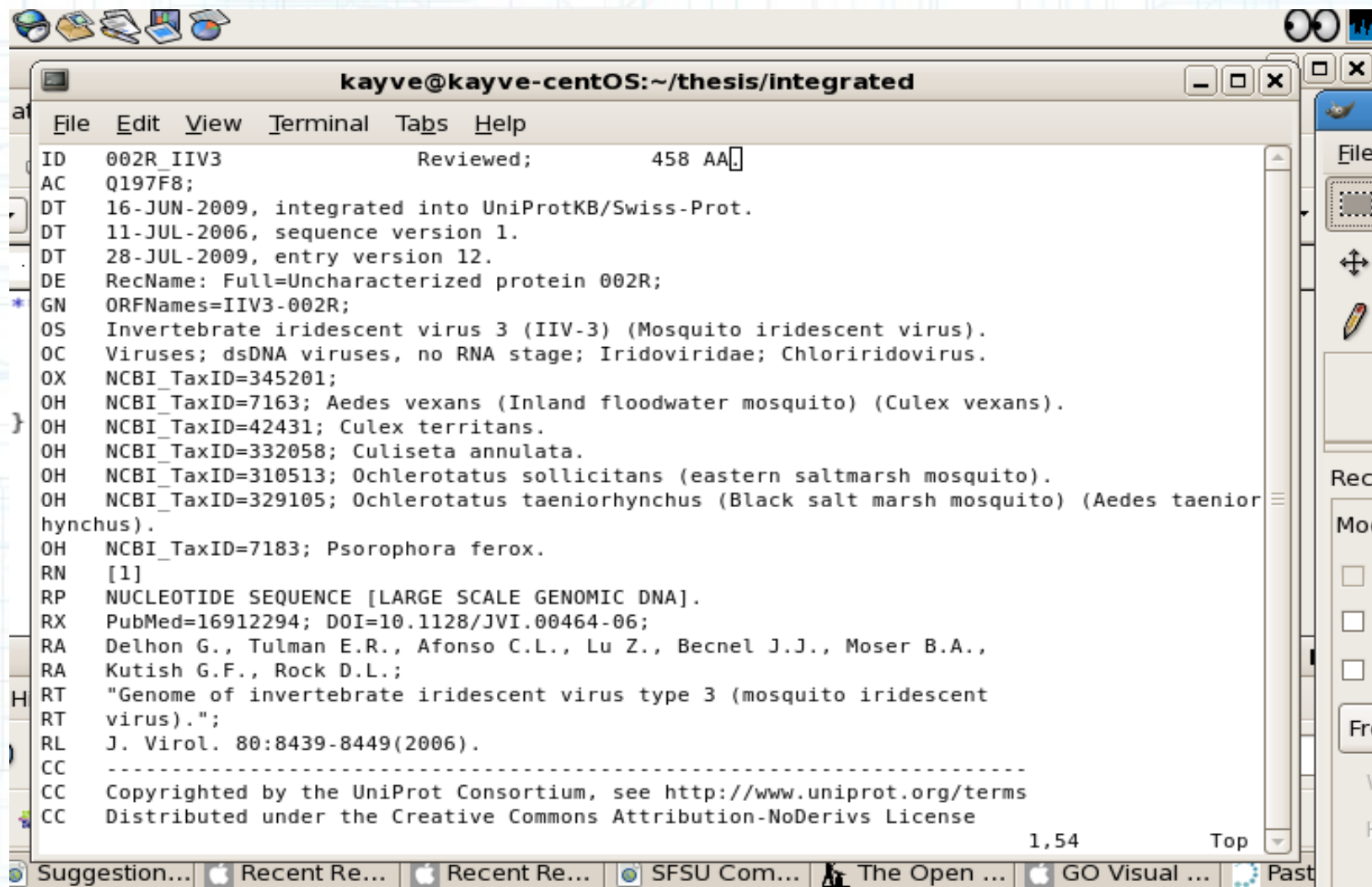
C Language Chosen Over Perl Due to Improved Performance

98			
99	Perl & C comparisons		
100			
101		MEAN	STDEV
102	<u>CentOS perl</u>	37.7	0.95
103	<u>CentOS C</u>	16.44	0.06
104	<u>Gentoo perl</u>	60	3.20
105	<u>Getnoo C</u>	31.51	1.99
106	<u>Ubuntu perl</u>	60.4	4.20
107	<u>Ubuntu C</u>	20.56	0.49
108			
109			
110			
111			
112			
113			
114			
115			
116			



- Three Linux distributions tested
- Each system tested both on C and Perl
- CentOS C chosen as best option

Structure of UniProt Flatfile Data



The screenshot shows a terminal window titled "kayve@kayve-centOS:~/thesis/integrated". The window displays the UniProt flatfile data for protein 002R_IIV3. The data is organized into sections: ID, AC, DT, DE, GN, OS, OC, OX, OH, RN, RP, RX, RA, RT, RL, CC, and a footer. The protein is identified as Invertebrate iridescent virus 3 (IIV-3) (Mosquito iridescent virus). The sequence is 458 AA long. The entry is reviewed. The data is copyrighted by the UniProt Consortium and distributed under the Creative Commons Attribution-NoDerivs License.

```
ID 002R_IIV3 Reviewed; 458 AA
AC Q197F8;
DT 16-JUN-2009, integrated into UniProtKB/Swiss-Prot.
DT 11-JUL-2006, sequence version 1.
DT 28-JUL-2009, entry version 12.
DE RecName: Full=Uncharacterized protein 002R;
GN ORFNames=IIV3-002R;
OS Invertebrate iridescent virus 3 (IIV-3) (Mosquito iridescent virus).
OC Viruses; dsDNA viruses, no RNA stage; Iridoviridae; Chloriridovirus.
OX NCBI_TaxID=345201;
OH NCBI_TaxID=7163; Aedes vexans (Inland floodwater mosquito) (Culex vexans).
OH NCBI_TaxID=42431; Culex territans.
OH NCBI_TaxID=332058; Culiseta annulata.
OH NCBI_TaxID=310513; Ochlerotatus sollicitans (eastern saltmarsh mosquito).
OH NCBI_TaxID=329105; Ochlerotatus taeniorhynchus (Black salt marsh mosquito) (Aedes taeniorhynchus).
OH NCBI_TaxID=7183; Psorophora ferox.
RN [1]
RP NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RX PubMed=16912294; DOI=10.1128/JVI.00464-06;
RA Delhon G., Tulman E.R., Afonso C.L., Lu Z., Becnel J.J., Moser B.A.,
RA Kutish G.F., Rock D.L.;
RT "Genome of invertebrate iridescent virus type 3 (mosquito iridescent
RT virus).";
RL J. Virol. 80:8439-8449(2006).
CC -----
CC Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC Distributed under the Creative Commons Attribution-NoDerivs License
```

1,54 Top

Suggestion... Recent Re... Recent Re... SFSU Com... The Open ... GO Visual ... Past

UniProt Line Types Utilized

- CC Line – Subcellular Location Comment Blocks

```

RE 5. V1101. 00.0433-0443(2000).
CC  -!- SUBCELLULAR LOCATION: Host membrane; Single-pass membrane protein
CC      (Potential).
CC  -----
CC  Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC  Distributed under the Creative Commons Attribution-NoDerivs License
CC  -----
DR  EMBL: D06433.92. AB02067.1. -. Genomic DNA
  
```

- DR Line – Database Cross Reference with Gene Ontology (GO) 7 digit codes

```

ot DR  GeneID; 1733039; -.
DR  GO; GO:0016021; C:integral to membrane; IEA:UniProtKB-KW.
PF  4: Predicted:
  
```

- FT Line - Feature Table Fixed Format Key Names

```

FI  /FTID=PRO_0000377966.
FT  TRANSMEM      4      24      Helical; (Potential).
FT  COMPRTAS      30     49      Pro-rich.
  
```

ProMog.c Module – Proteomic DeMographics

- Fixed format data extraction from FT lines
- Regular Expressions (RegEx) in C used
- CC line subcellular location (SCL) data RegEx'ed
- Database Cross Reference (DR line) RegEx'ed
- Gene Ontology (GO) data contained on DR Lines
- Ahram et al signal peptide removal followed
- Total of 59 RegEx key words employed

RegEx in C – the regex_t Structure Definition

```
00164 /* the biggie, a compiled RE (or rather, a front end to same) */
00165 typedef struct {
00166     int re_magic;           /* magic number */
00167     size_t re_nsub;         /* number of subexpressions */
00168     long re_info;           /* information about RE */
00169 #       define REG_UBACKREF      000001
00170 #       define REG_ULOOKAHEAD    000002
00171 #       define REG_UBOUNDS      000004
00172 #       define REG_UBRACES      000010
00173 #       define REG_UBSALNUM     000020
00174 #       define REG_UPBOTCH      000040
00175 #       define REG_UBBS         000100
00176 #       define REG_UNONPOSIX    000200
00177 #       define REG_UUNSPEC      000400
00178 #       define REG_UUNPORT      001000
00179 #       define REG_ULOCALE      002000
00180 #       define REG_UEMPTYMATCH  004000
00181 #       define REG_UIMPOSSIBLE  010000
00182 #       define REG_USHORTEST    020000
00183     int re_csize;           /* sizeof(character) */
00184     char *re_endp;           /* backward compatibility kludge */
00185     /* the rest is opaque pointers to hidden innards */
00186     char *re_guts;           /* `char *' is more portable than `void *' */
00187     char *re_fns;
00188 } regex_t;
```

ProMog.C RegEx Implementation

- Declarations of regex_t Arrays

```
/******  
 * REGular EXpressions compiled variables  
******/  
regex_t rgx_array[REGEX_COUNT], rgx_GO_array[GO_COUNT],  
    rgx_GO_minor_array[GO_MINOR_COUNT];  
int regex_status;
```

- Precomplitation of regex_t Structures

```
/******  
 * REGular EXpression COMPilations  
******/  
for (i=0;i<REGEX_COUNT;i++) {  
    regex_status = regcomp(&rgx_array[i],REGEX_RAW_ARRAY[i] , REG_EXTENDED|REG_NOSUB);  
    if (regex_status) {  
        fprintf(stderr, "Could not compile regex for %s\n",REGEX_RAW_ARRAY[i]);  
        exit(REGEX_ERR);  
    }  
}
```

- Execution of Precompiled regex_t Structures

```
for(i=0;i<REGEX_COUNT;i++) {  
    if (!(regexec(&rgx_array[i], line, (size_t)0,NULL,0))) {  
        is_SCL_ARRAY[i] = TRUE;  
        is_REMAINDER = FALSE;  
    }  
}
```


Boolean Flags and Their Tabulators

- Boolean flag `is_TRANSMEM` detects fixed format key name

```
if ((block[line_begin] == 'F') && (block[line_begin+1] == 'T')) {  
    /*****  
    * Feature Table (FT) line  
    * http://www.expasy.org/sprot/userman.html#FT\_line  
    *  
    *****/  
    if ((block[line_begin+5] == 'T') && (block[line_begin+6] == 'R') &&  
        (block[line_begin+7] == 'A') && (block[line_begin+8] == 'N') &&  
        (block[line_begin+9] == 'S') && (block[line_begin+10] == 'M') &&  
        (block[line_begin+11] == 'E') && (block[line_begin+12] == 'M')) {  
        is_FT_TRANSMEM = TRUE;  
        is_REMAINDER = FALSE;  
    } /--- if FT TRANSMEM ----//
```

- Paired tabulator `hum_transmem` increments at end of record

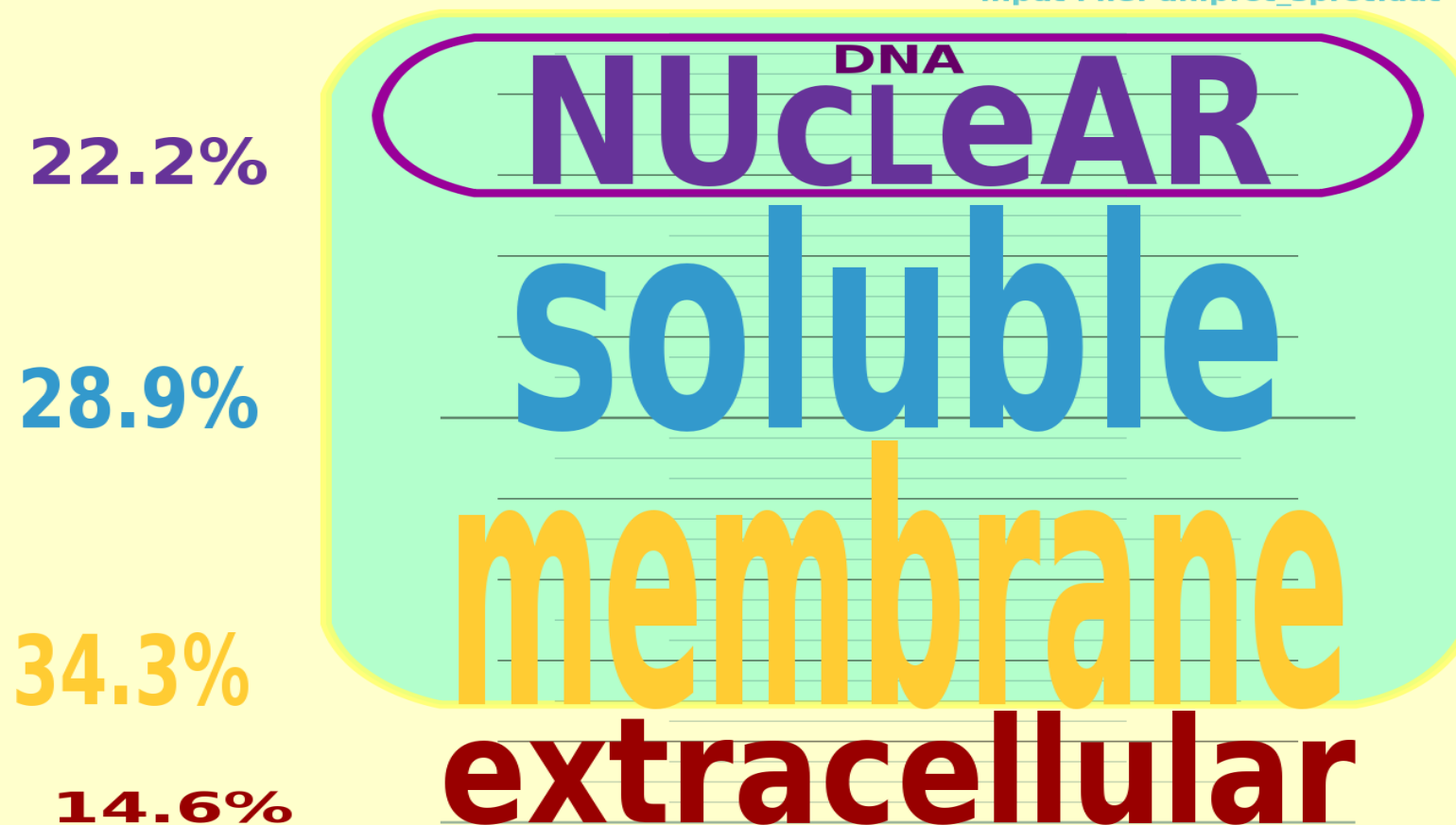
```
if ((block[line_begin] == '/') && (block[line_begin+1] == '/')) {  
    /*****  
    * END OF RECORD  
    *****/  
    tot_proteins++;  
    if (n_prot_lines > max_prot_lines)  
        max_prot_chars = this_prot_chars;  
    if (this_is_human) {  
        /*****  
        * HUMAN DATA  
        *****/  
        if (has_SCI) {  
            hum_REMAINDER++;  
            if (is_FT_TRANSMEM)  
                hum_transmem++;  
            if (is_FT_INTRAMEM)
```

Cellgram.c Module – Bio-Relevant Diagrams

May 4, 2011 23:53

brain

Input File: uniprot_sprot.dat



Cairo Graphics



- C Language Library with bindings to other Programming Languages
- Distributed under Gnu LGPL or Mozilla MPL Licenses
- Integrated with GTK as part of the Gnome Desktop
- Used to produce the png Cellgram images
- Online recipes used for foundational implementation

Cairo Rounded Rectangle



The screenshot shows a web browser window displaying the Cairo website. The address bar shows the URL `http://cairographics.org/samples/rounded_rectangle/`. The page features the Cairo logo, which consists of two orange beetles flanking the word "cairo" in a blue serif font. Below the logo is a dark blue navigation bar with links for "News", "Download", "Documentation", "Contact", and "Examples". The "rounded rectangle" section is highlighted in orange. It contains a code block with C code for creating a rounded rectangle using the Cairo library, and a visual representation of the resulting rounded rectangle.

```
/* a custom shape that could be wrapped in a function */
double x      = 25.6,      /* parameters like cairo_rectangle */
y      = 25.6,
width    = 204.8,
height   = 204.8,
aspect   = 1.0,          /* aspect ratio */
corner_radius = height / 10.0; /* and corner curvature radius */

double radius = corner_radius / aspect;
double degrees = M_PI / 180.0;

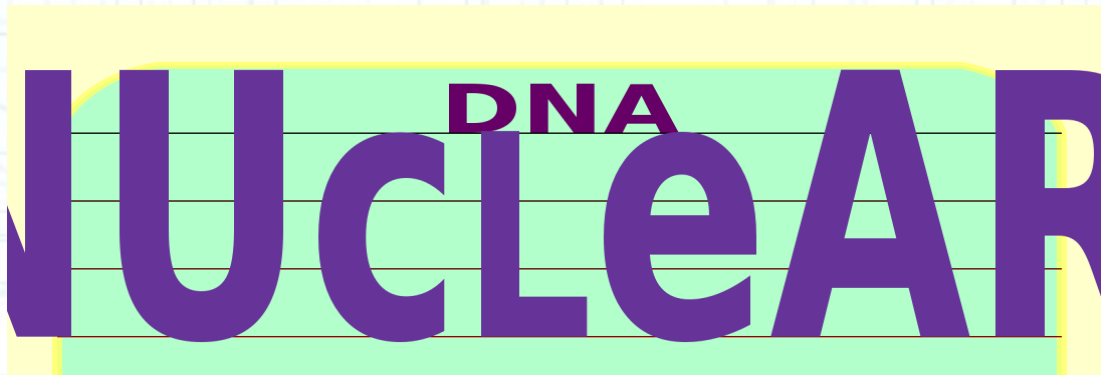
cairo_new_sub_path (cr);
cairo_arc (cr, x + width - radius, y + radius, radius, -90 * degrees, 0 * degrees);
cairo_arc (cr, x + width - radius, y + height - radius, radius, 0 * degrees, 90 * degrees);
cairo_arc (cr, x + radius, y + height - radius, radius, 90 * degrees, 180 * degrees);
```



- Rounded rectangle recipe served as “plasma membrane”
- Robust online documentation and e-mailing list support

Resizing Word Widths to Fit Cellgram Constraints

- Early experiment with changing word height



- Solution: “Squish ratio”

$$e^{(A^2z + Ax + y)} = \text{squish ratio}$$

A: Real width of text

x,y,z: three unknowns

Procedure: Choose three values for A, and respective “squish ratios,” and solve system of equations.

Wolfram Alpha Online Math Service



solve $\exp(160000z + 400x + y) = 1$ and $\exp(4000000z + 2000x + y) = 0.4$ and $\exp(56250000z + 7500x + y) = 0.133$

Input interpretation:

solve	$\exp(160\,000z + 400x + y) = 1$
	$\exp(4\,000\,000z + 2000x + y) = 0.4$
	$\exp(56\,250\,000z + 7500x + y) = 0.133$

Result:

$$z = \frac{1}{624\,800\,000} (110 i \pi c_1 - 142 i \pi c_2 + 32 i \pi c_3 - 16 (3 \log(2) + 3 \log(5) - \log(7) - \log(19)) - 71 (\log(2) - \log(5)))$$

$$\text{and } x = \frac{1}{6\,248\,000} (-10\,450 i \pi c_1 + 11\,218 i \pi c_2 - 768 i \pi c_3 + 384 (3 \log(2) + 3 \log(5) - \log(7) - \log(19)) + 5609 (\log(2) - \log(5)))$$

- Online resource at <http://www.wolframalpha.com/>
- Produces pdf output file of results

Resizing Text with `cairo_scale()`

- Solutions loaded in the const ints, some inverted

```
const double SQUISH_SQRD = 24067359.9345;  
const double SQUISH_COEFF = 1668.29167;  
const double SQUISH_TERM = 0.118221111688; /**/  
----- double SQUISH_COEFF = 13101100.7122402.
```

- Variable `tx_squish` calculated with exponential function

```
tx_squish = exp(tx_width*tx_width/SQUISH_SQRD-1*(tx_width/SQUISH_COEFF)+  
SQUISH_TERM);
```

- `cairo_scale()` used to implement scaling
- Geometric scaling undone division by `txsq_DNA`

```
cairo_move_to (cr, txo_x, txo_y);  
cairo_scale(cr, tx_squish/txsq_DNA, 1);  
cairo_show_text (cr, "NUc");
```

Total Proteins in Manually Annotated May 3 Release

The image shows a web browser window displaying the UniProt release notes for May 3, 2011. The browser's address bar shows the URL `ftp://ftp.uniprot.org/pub/databases/uniprot/relnotes.txt`. The page content includes the title "UniProt Release 2011_05" and a paragraph stating that the UniProt consortium (EBI, SIB, and PIR) is announcing the UniProt Knowledgebase (UniProtKB) Release 2011_05 (03-May-2011). It also provides the entry counts for UniProtKB Release 2011_05: 15,590,885 total entries, with 528,048 from Swiss-Prot and 15,062,837 from TrEMBL. Additionally, it mentions that UniRef100 Release 2011_05 consists of 12,831,896 entries.

Below the browser window, a terminal window is open, showing the command prompt `kayve@kayve-centOS:~/thesis/integrated`. The terminal output displays the following protein counts:

```
-----  
There are 528048 total proteins  
total FT TRANSMEM proteins: 70729  
total FT INTRAMEM proteins: 807  
total proteins with covalent lipid binding: 6937
```

Verification of correctness of protein count (528,048 total proteins)

Proteins With Relevant Annotations

```
o: total proteins with Gene Ontology DNA binding : 29707
FY total proteins with no CC SUBCELLULAR LOCATION annotation: 222196
total total membrane proteins: 66863
total cytoplasmic proteins: 160998
total extracellular proteins: 45352
total nuclear proteins: 58603
REMAINDER total proteins: 100
-----
total brain proteins: 16119
brain nuclear proteins: 4015
brain cytoplasmic proteins: 5223
brain membrane proteins: 6190
brain extracellular proteins: 2629
-----
06 total muscle proteins: 833
muscle nuclear proteins: 215
muscle cytoplasmic proteins: 273
muscle membrane proteins: 292
muscle extracellular proteins: 129
-----
The protein with the most lines has 6304 lines
```

Total proteins used in Cellgram calculation – 331,819 or 62.8% of the 528,048

Human Proteins With Relevant Annotations

```
5 there are a total of 43430000 lines
6 -----HUMAN PROTEINS-----
7 human proteins: 20239
8 human ET TRANSMEM proteins: 5148

9 human proteins with no CC SUBCELLULAR LOCAT
0human total membrane proteins: 3434
1human cytoplasmic proteins: 5369
2human extracellular proteins: 3767
3human nuclear proteins: 5140
4REMAINDER human proteins: 9
5-----
6 There are 528048 total proteins
```

- Relevant human proteins totaled 17,710
– equivalent to 87.5%
- Higher rate likely due to importance of human proteins in medical research

Human Protein Manually Annotated Database Cellgram

May 16, 2011 9:17

human

Input File: uniprot_sprot.dat

29.0%

^{DNA}
NUCLeAR

30.3%

soluble

19.4%

membrane

21.3%

extracellular

Summary

- FMM/TMP topology/compartmental theory important to medicine
- UniProt Release of TrEMBL data undergoing exponential growth
- C language provides RegEx & superior performance to Perl
- C language provides support for regular expressions (RegEx)
- ProMog.c extracts demographic protein data from UniProt files
- Cellgram.c produces SCL diagrams using Cairo Graphics
- Wolfram Alpha used to resize words horizontally
- 62.8% of total, and 87.5% of human proteins found relevant