ProMog CellGrams

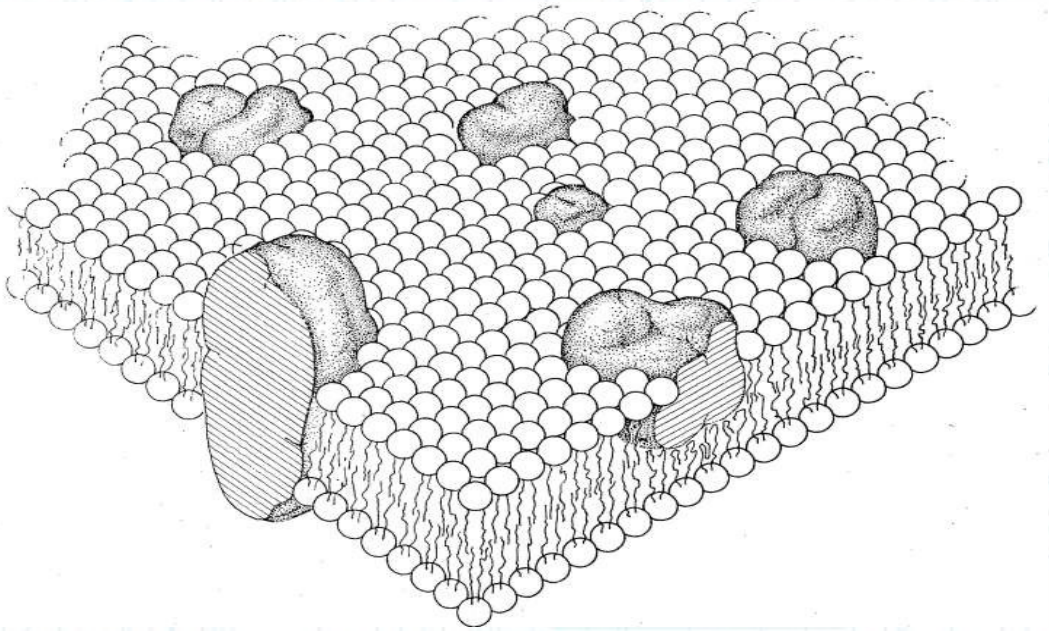
Evidence-Based Proteome-Wide Subcellular Location – A Basis Partitioning for Classification Systems

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

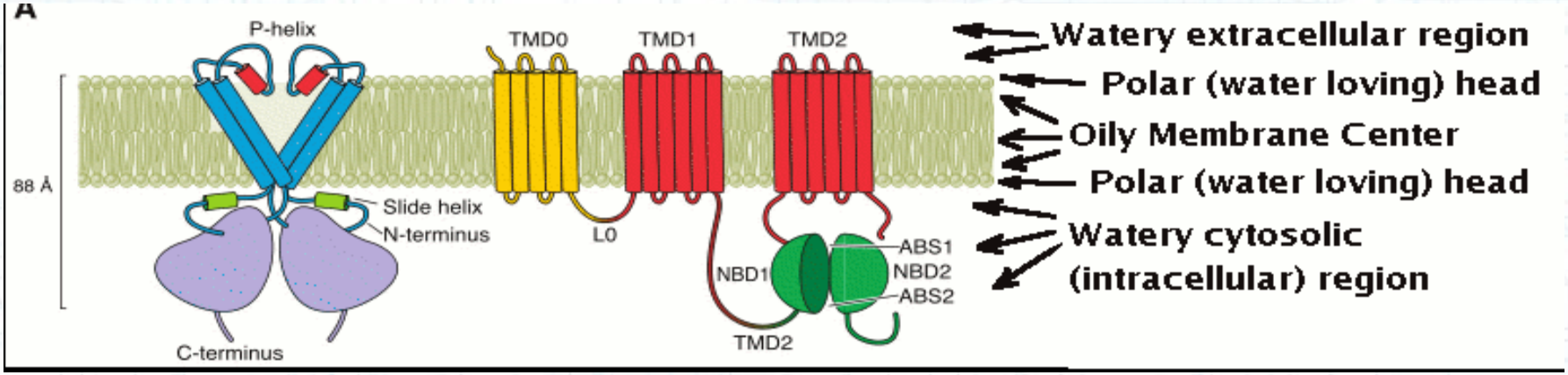
Fluid Mosaic Model (FMM)



Proposed by Singer and Nicholson in 1972

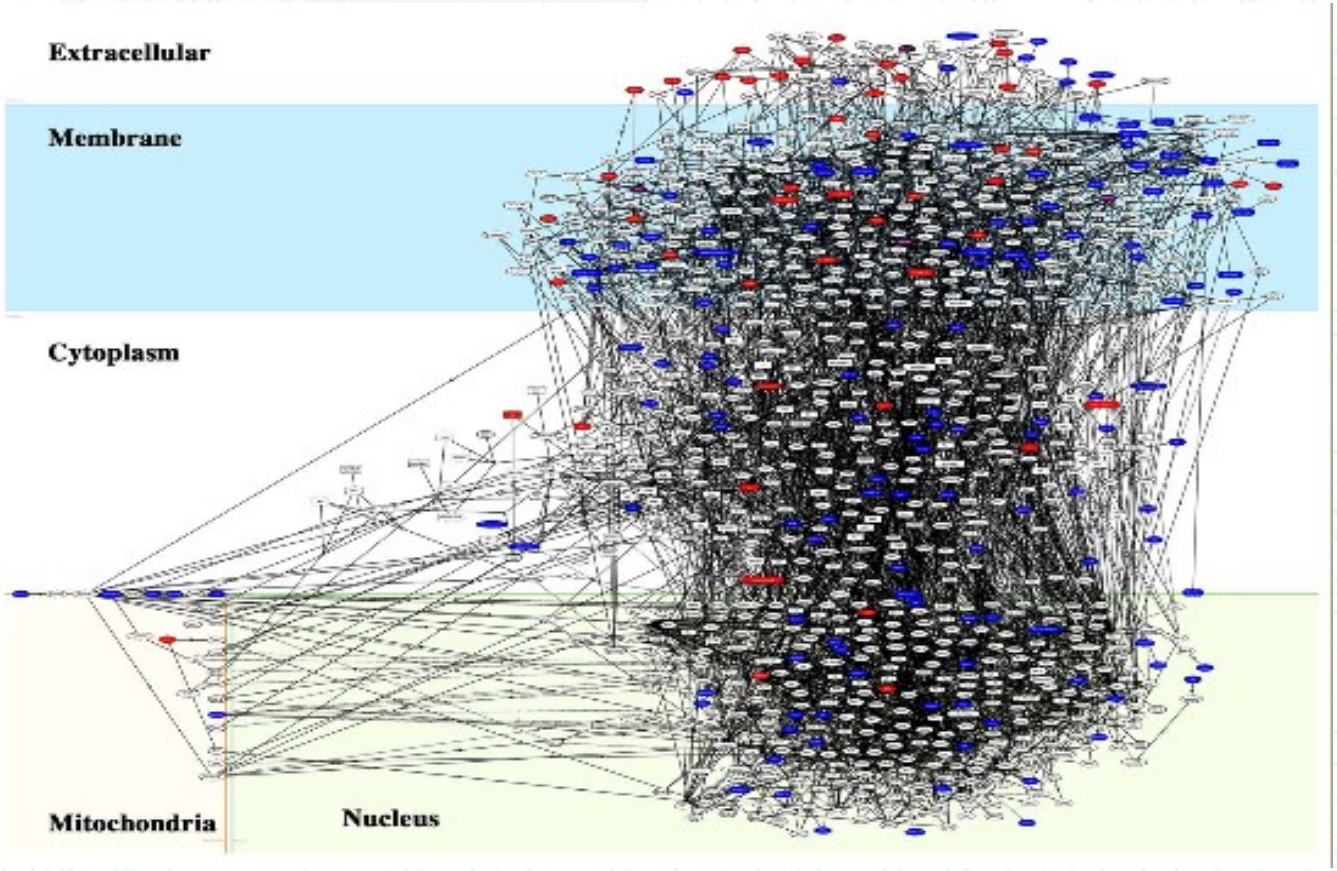
- Theory states some proteins are lodged in membranes
- Oily membranes separate watery compartments
- Special proteins constrained in psuedo-2D
- Serve as gatekeepers between cellular regions

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

FMM is Integral to Modern Molecular Theory

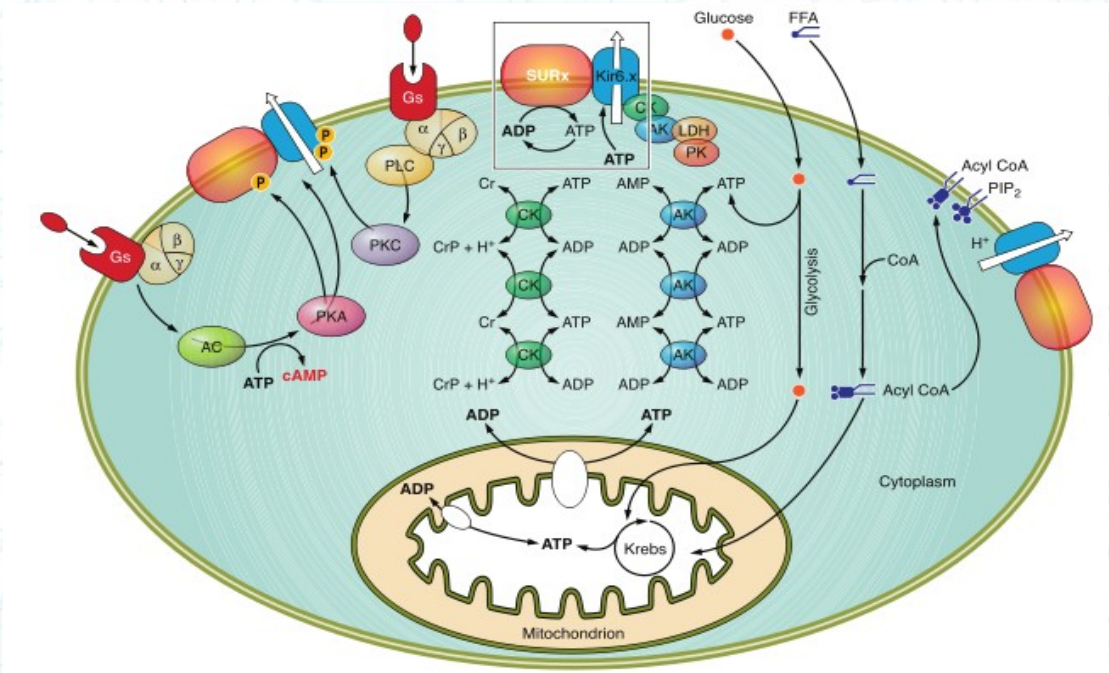


FIG. 2. Complexities of K_{ATP} channel regulation in striated muscle. The balance of ATP synthesis and usage, reflected by ATP and ADP levels, is the major direct determinant of channel activity (box). Metabolic enzymes, including adenylate kinase (AK), creatine kinase (CK), and lactate dehydrogenase (LDH), in the cytoplasm and physically associated with the channel may serve to amplify metabolic changes, or locally buffer and control ATP/ADP levels, thereby fine-tuning channel activity. Nonnucleotide ligands, including PIP_2, acyl-CoA, and H^+, may also play a key role. PIP_2 and acyl-CoAs have powerful stimulatory effects that are antagonistic to ATP inhibition. In addition, hormone receptor activation can lead to protein phosphorylation (P) with both stimulatory and inhibitory effects on the channel. However, none of these molecules acts in isolation, and the resultant K_{ATP} channel activity is an integrated response to a myriad of interrelated metabolic signals.

Membranes and TMPs define biological function

TMP Topology

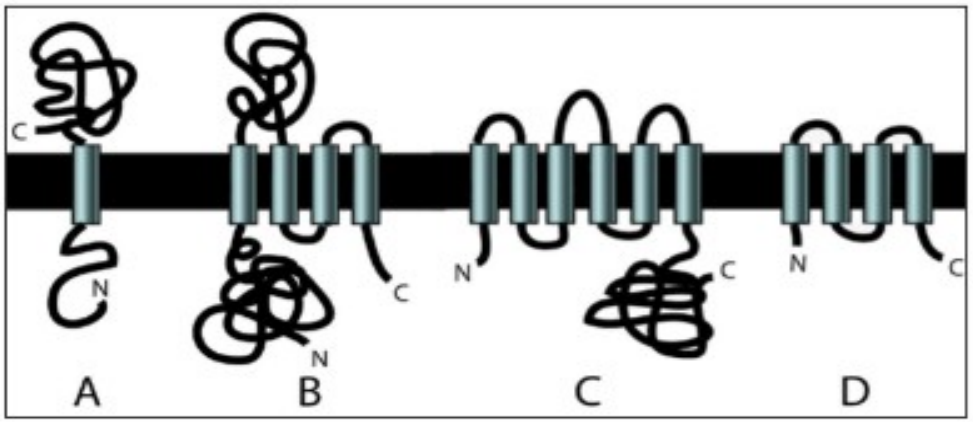


Figure 1 Schematic diagram of TM protein topology. The number of TM helices, the lengths of the N- and C-terminal domains, and the sizes of the interhelical loops or domains vary. Some TM proteins are located predominantly within the membrane (D), but others have more extensive N- or C-termini and/or cytoplasmic or periplasmic loops (A, B, C) that can form independently folding globular domains.

- 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

# of TMPs (% of total)
Topology
Prediction Methods

Protein Data Bank (PDB)—37,490 Gene products

TMHMM
6511 (17.4%)

SOSUI
8672 (23.1%)

HMMTOP
13296 (35.5%)

waveTM
16944 (45.2%)

TopPred
25751 (68.7%)

Discarded TopPred as unreliable

utilized SignalP to eliminate false positives

Final Estimated Results

TMHMM
5465 (14.6%)

SOSUI
7043 (18.8%)

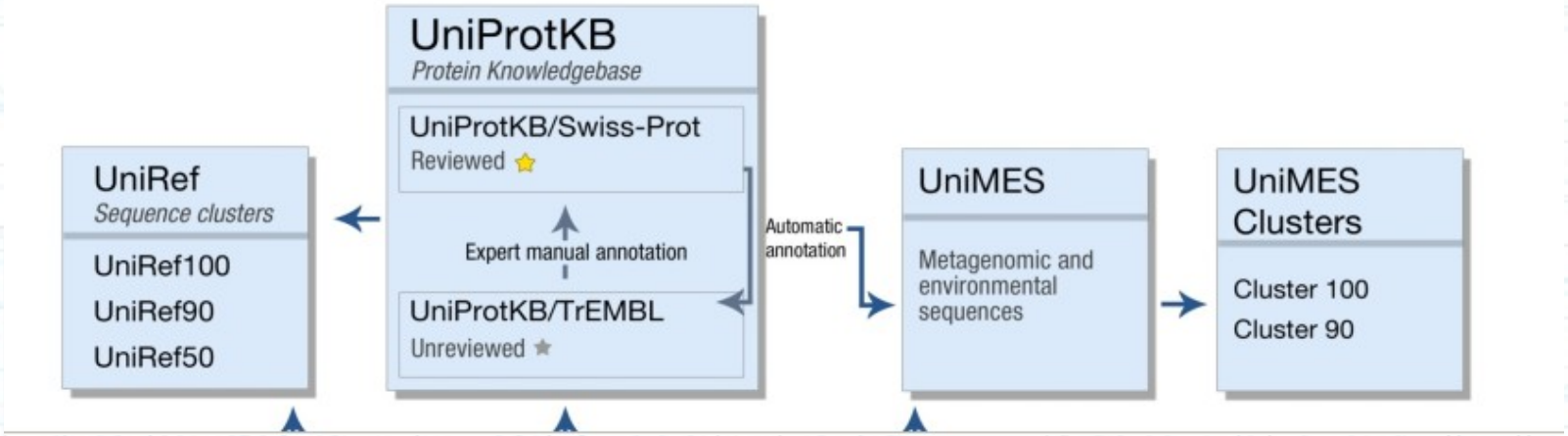
HMMTOP
3091 (23.3%)

waveTM
14678 (39.2%)

The Universal Protein Resource (UniProt)

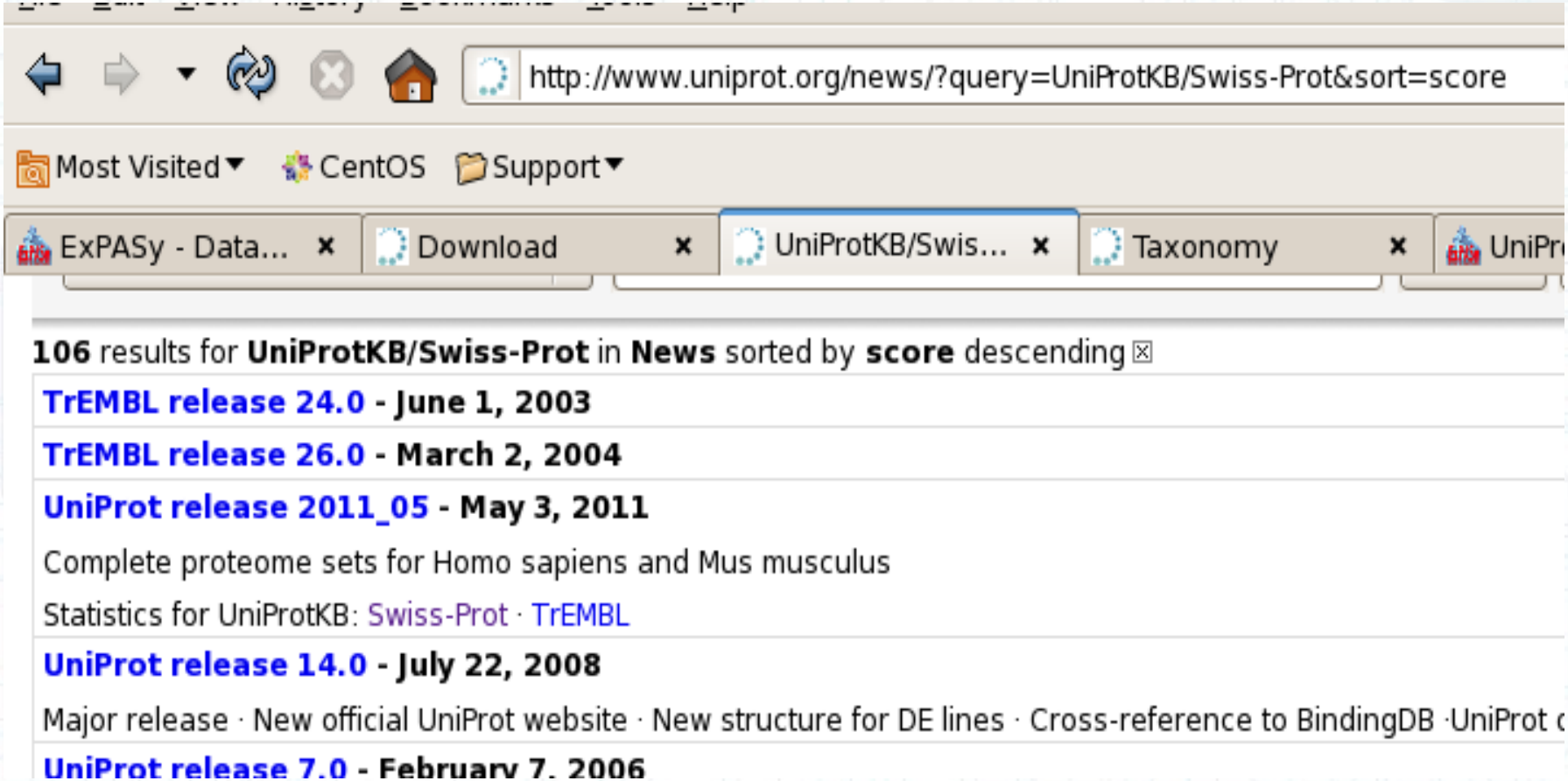
About UniProt

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



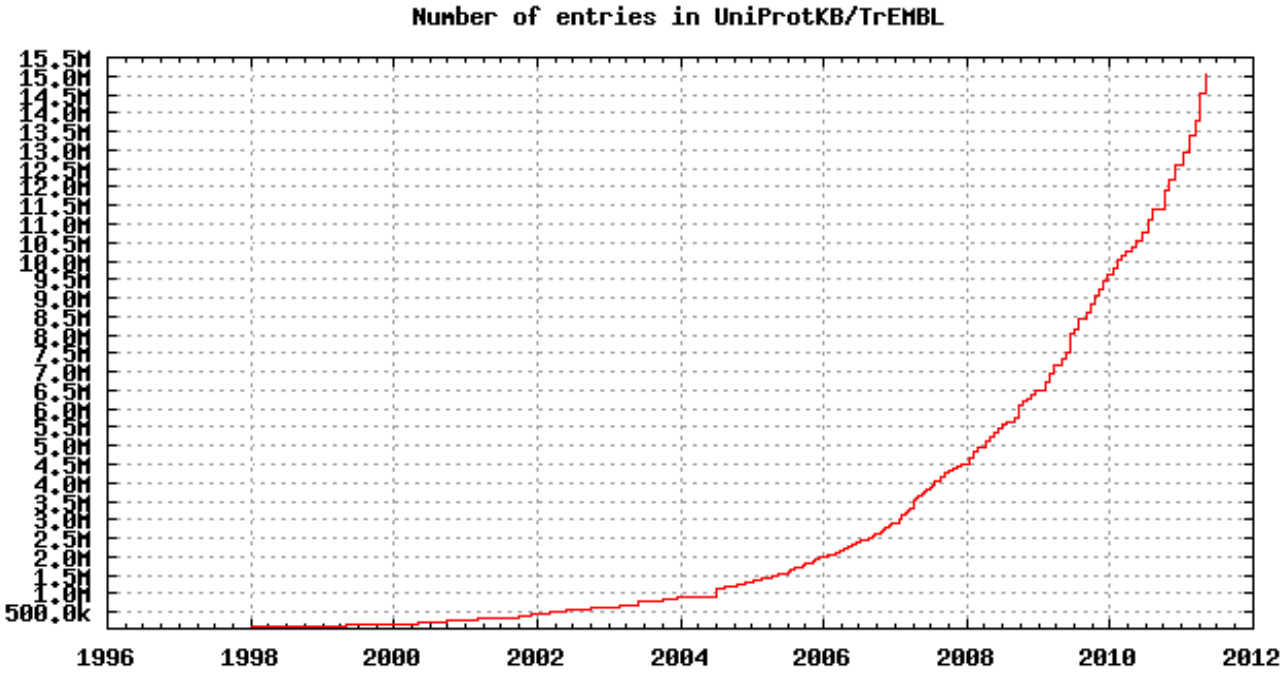
UniProt periodically updates release data

Exponential Growth of UniProt TrEMBL Data

EBI > Databases > Protein > UniProt > UniProtKB/TrEMBL

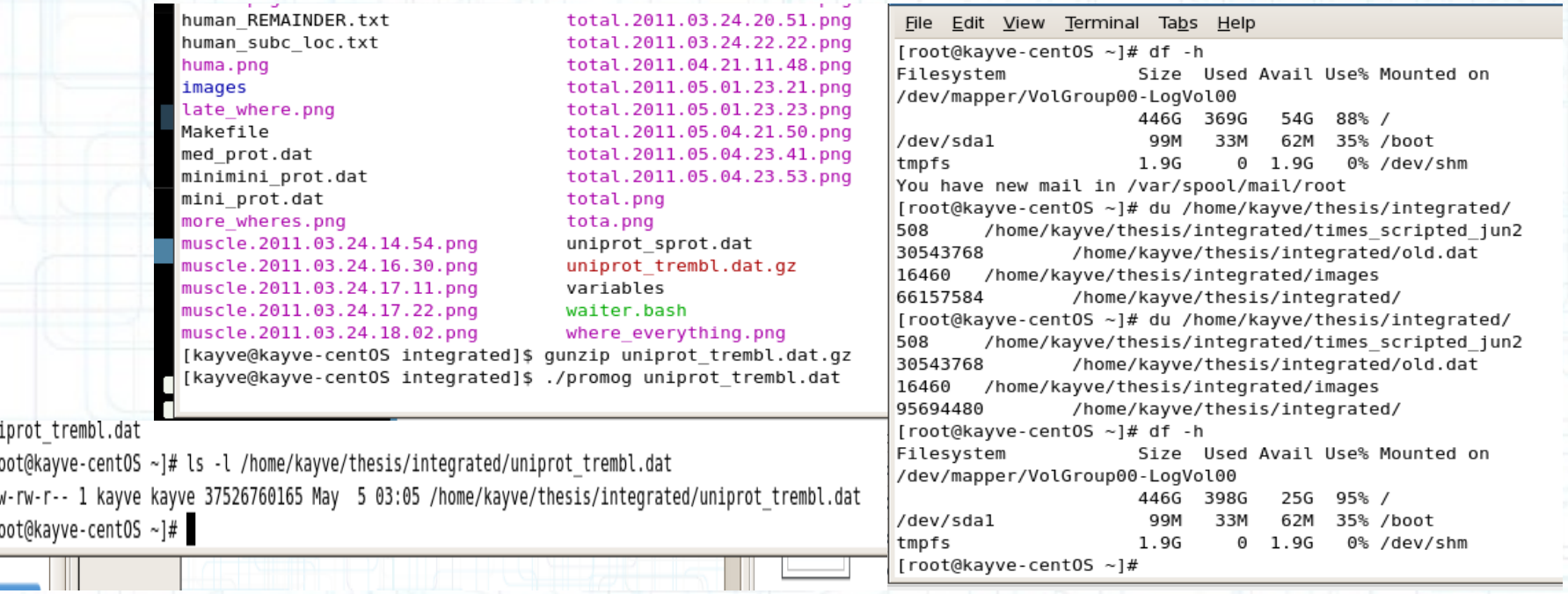
UniProtKB/TrEMBL - Current Release Statistics

The growth of the database is summarized below.



Over the past decade, data has exponentially increased

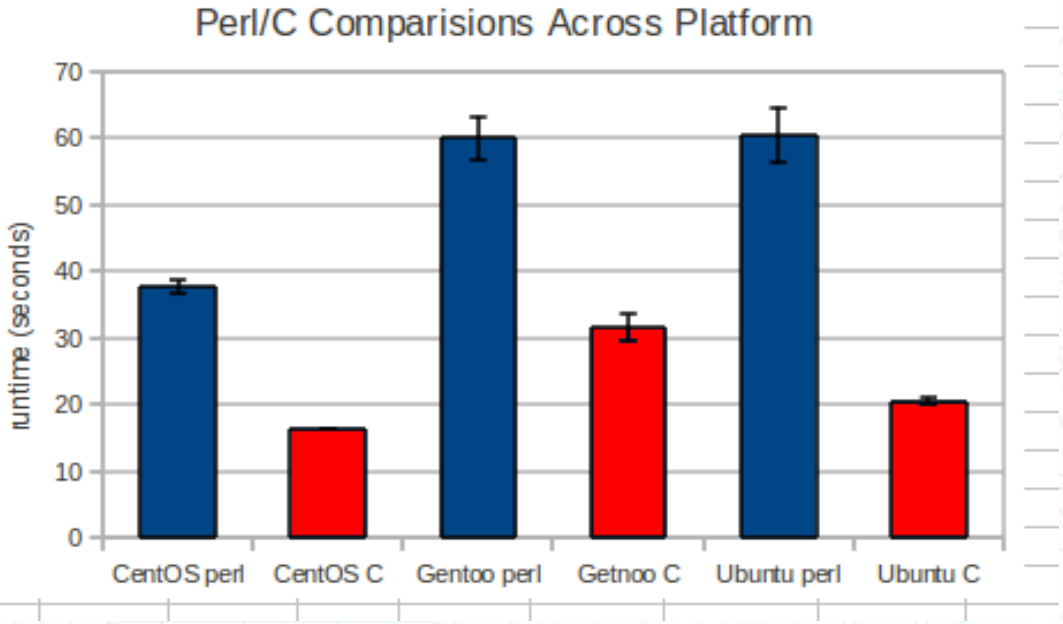
TrEMBL File Data Load



- 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

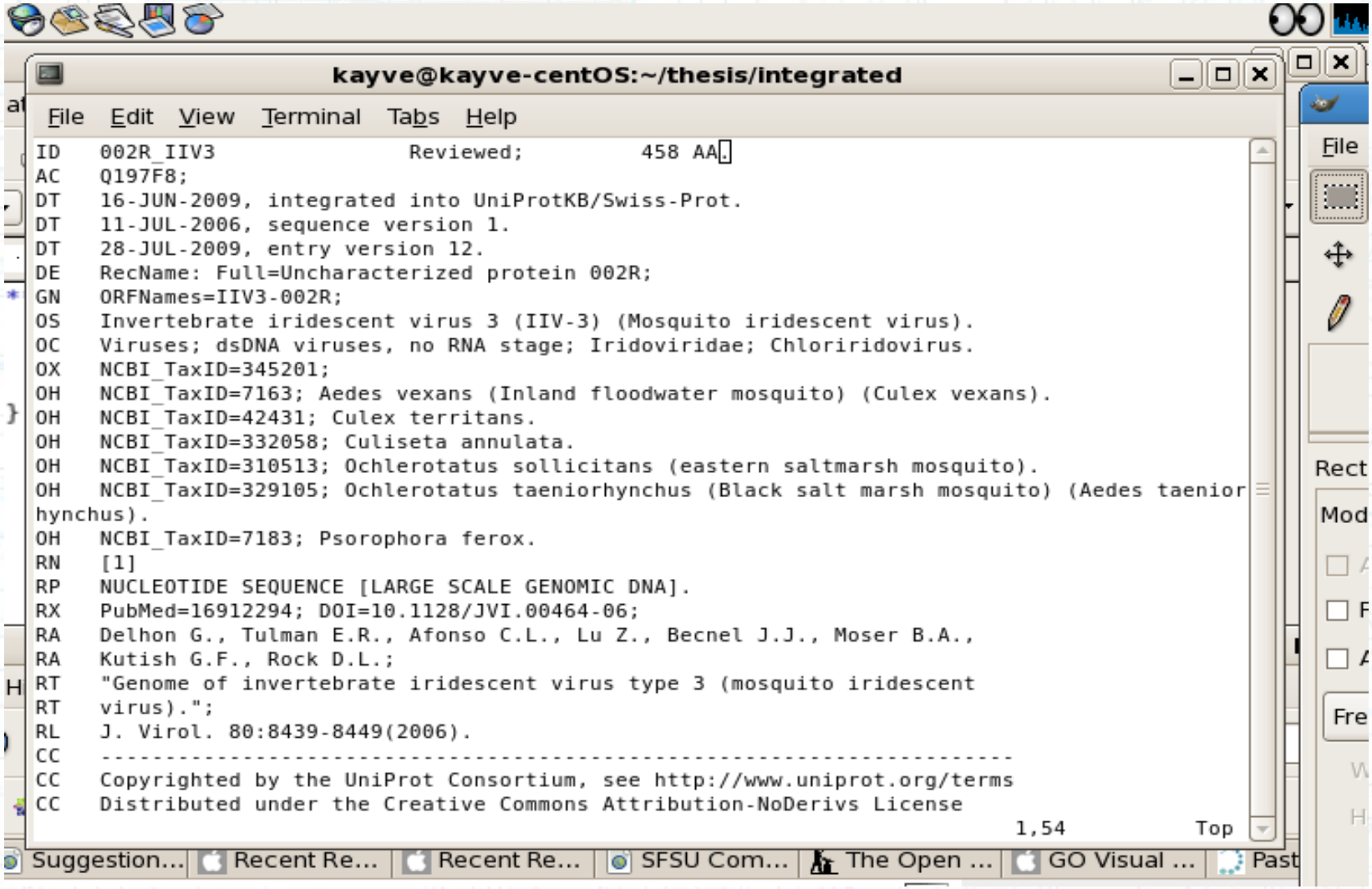
C Language Chosen Over Perl Due to Improved Performance

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



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

Structure of UniProt Flatfile Data



UniProt Line Types Utilized

- CC Line – Subcellular Location Comment Blocks

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

```
DR   GeneID; 1733059; -.
DR   GO; GO:0016021; C:integral to membrane; IEA:UniProtKB-KW.
PE   4: Predicted;
```

- FT Line - Feature Table Fixed Format Key Names

```
FT                                   /FTId=PRO_0000377966.
FT   TRANSMEM      4     24          Helical; (Potential).
FT   COMPBIAS     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

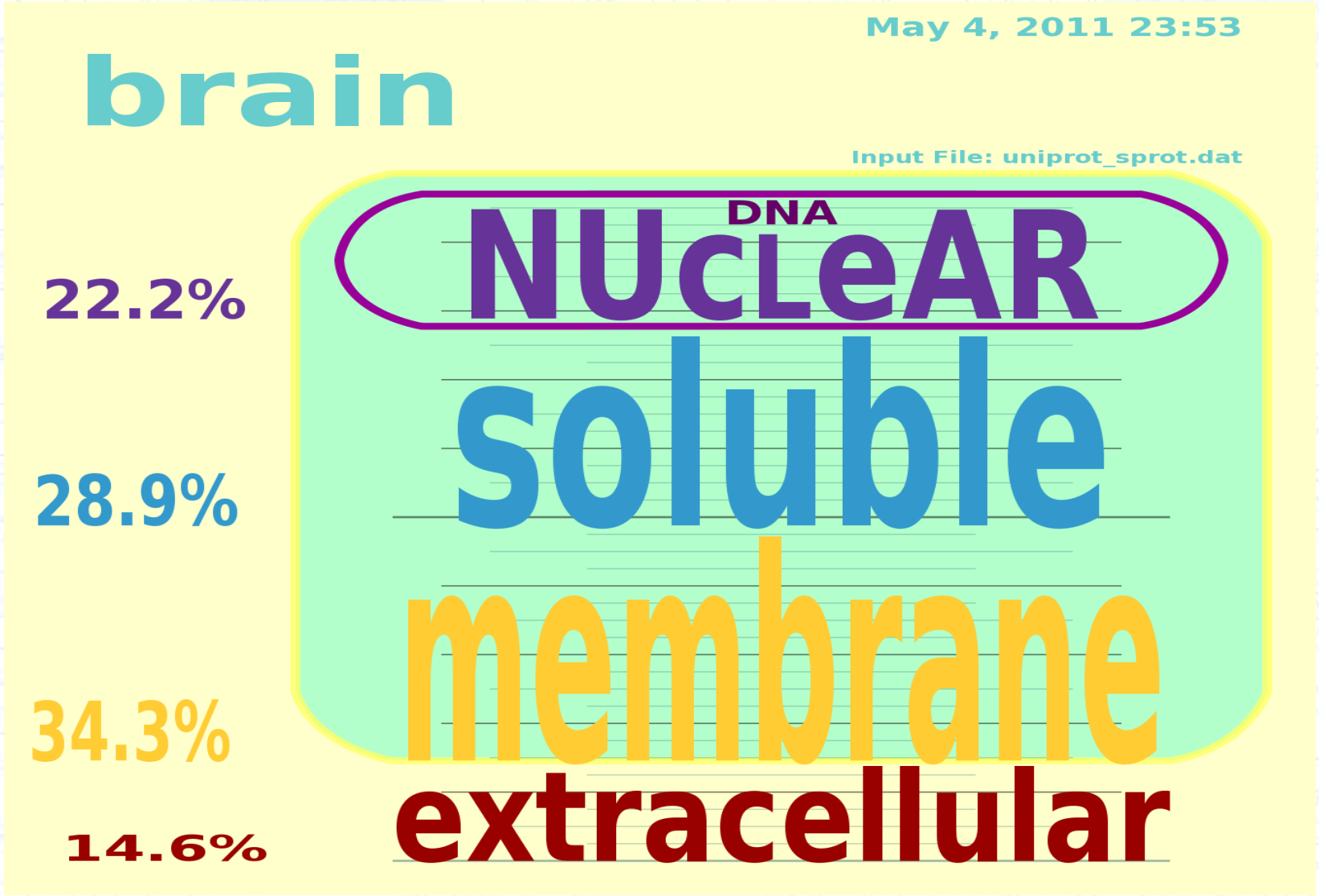
- Boolean flag is_TRANSMEM detects fixed format key name

```
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_SCL) {
                    hum_REMAINDER++;
                  if (is_FT_TRANSMEM)
                    hum_transmem++;
                  if (is_FT_INTRAMEM)
```

Cellgram.c Module – Bio-Relevant Diagrams



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

http://cairographics.org/samples/rounded_rectangle/

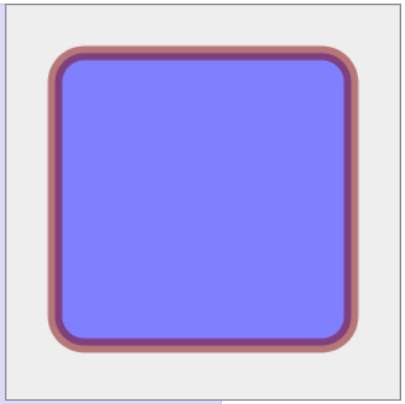
Most Visited CentOS Support

cairo

News | Download | Documentation | Contact | Examples

rounded rectangle

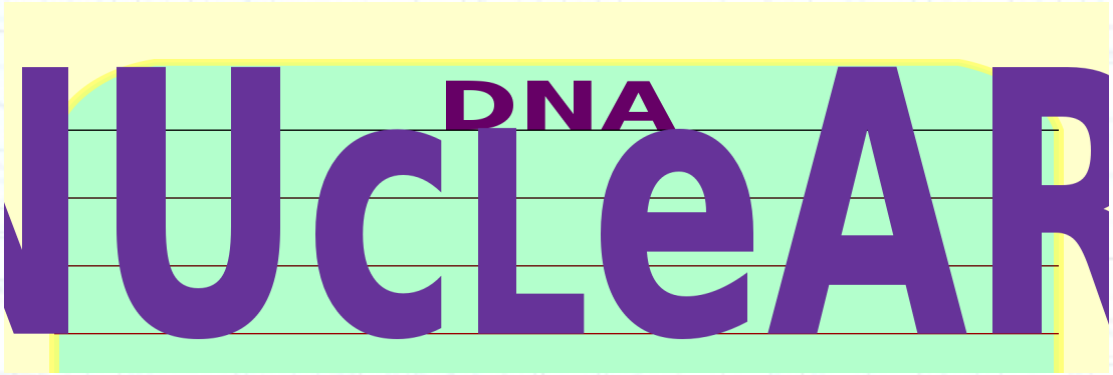
```
/* a custom shape that could be wrapped in a function */
double x         = 25.6,        /* parameters like cairo_rectangle */
       y         = 25.6,
       width         = 204.8,
       height        = 204.8,
       aspect        = 1.0,     /* aspect ratio */
       corner_radius = height / 10.0;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

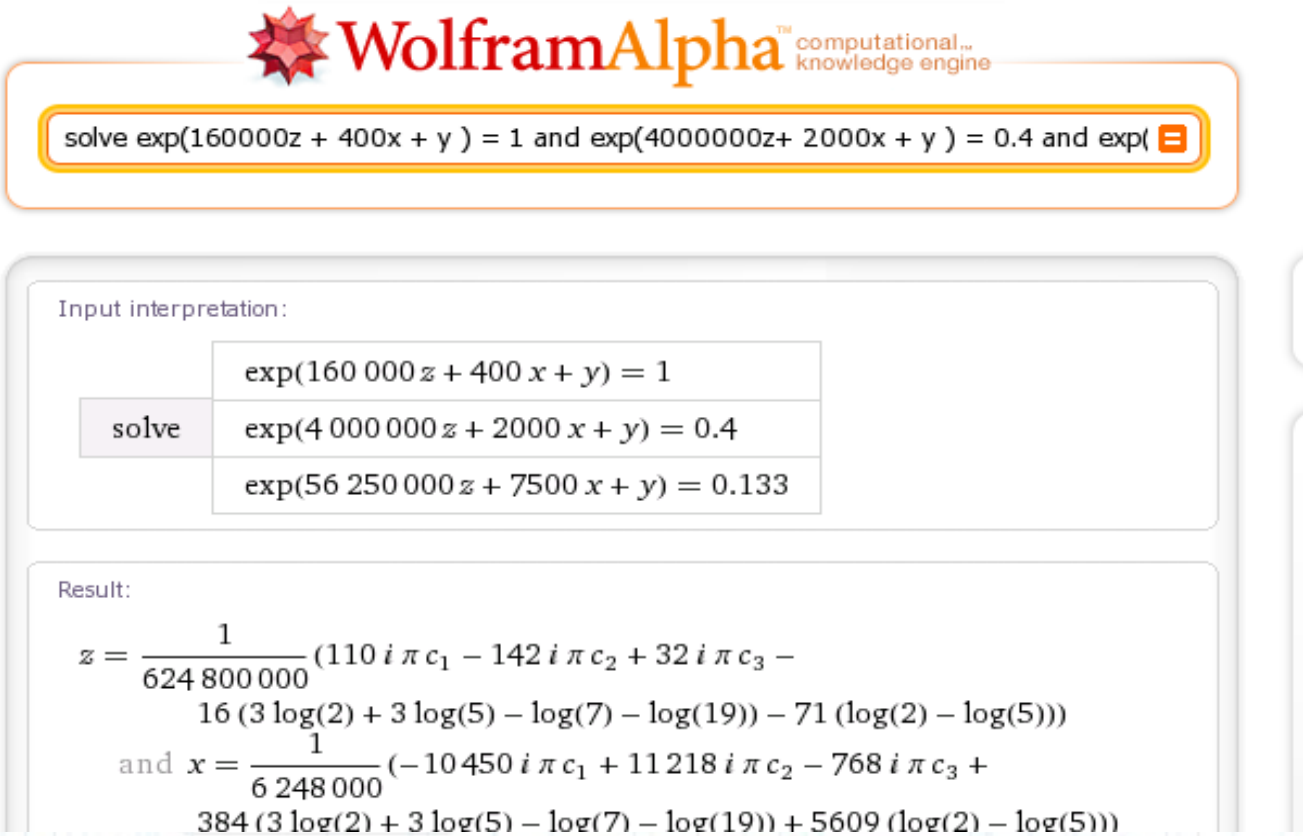


- Solution: "Squish ratio"

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

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



- 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; /**/
```

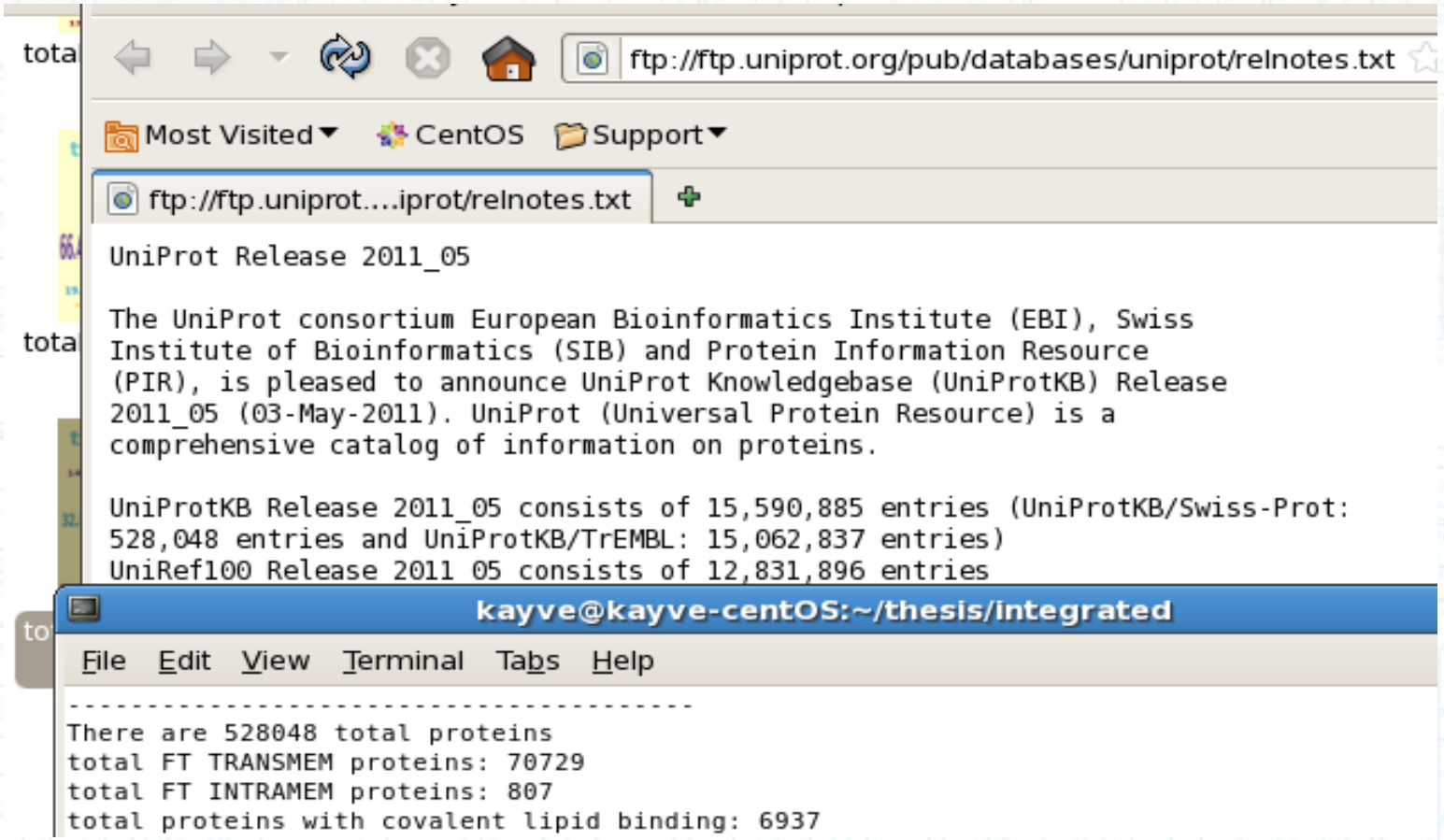
- Variable tx_squish calculated with exponential function

```
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



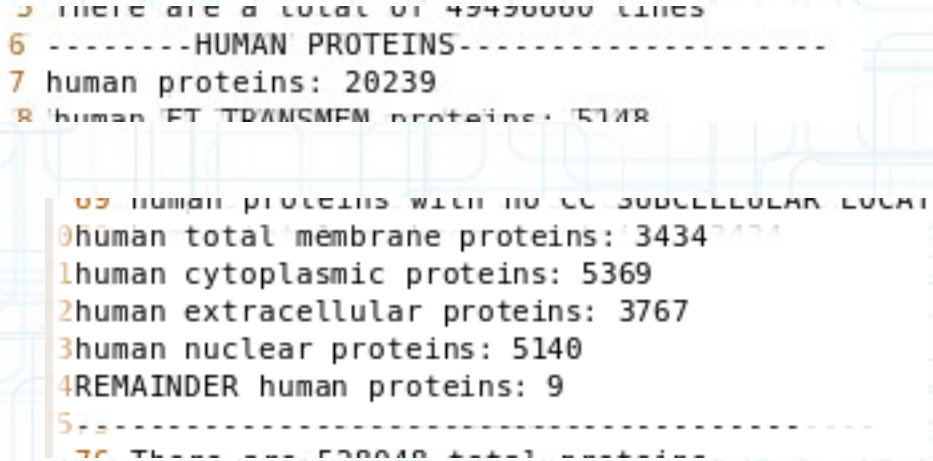
Verification of correctness of protein count (528,048 total proteins)

Proteins With Relevant Annotations

```
8: total proteins with Gene Ontology "DNA binding" : 29787
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
----------------------------------------
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
```

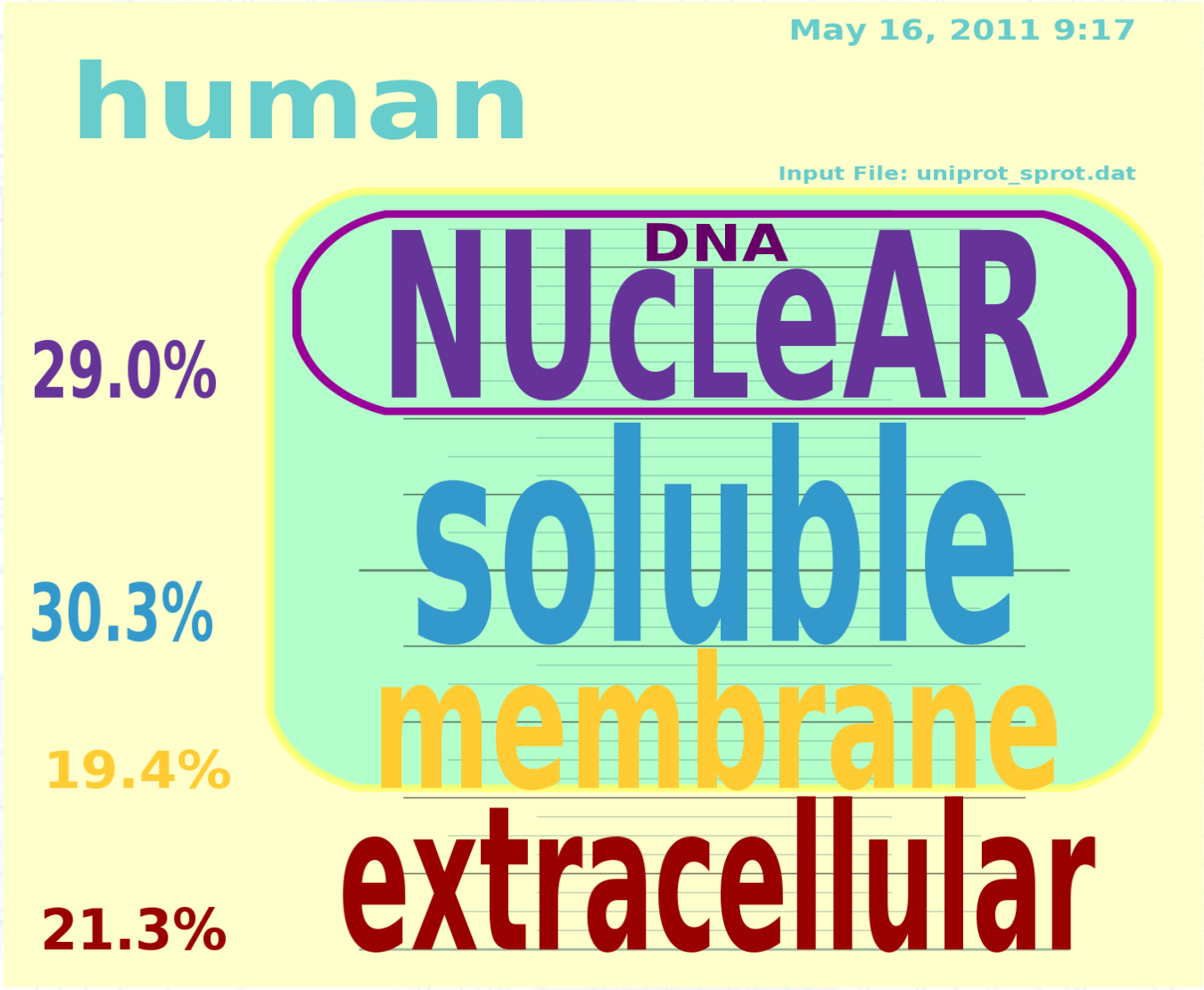
Total proteins used in Cellgram calculation – 331,819 or 62.8% of the 528,048

Human Proteins With Relevant Annotations



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

Human Protein Manually Annotated Database Cellgram



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 horizonally
- 62.8% of total, and 87.5% of human proteins found relevant