PatentInformatics

An Altoris, Inc. Project

Hugo O. Villar, Ph.D., MBA

Who We Are

San Diego based software development company

Contract Software Development

Customers in US, Europe and Japan.

CHEMAPPS™

SARvision™ PLUS

PatentInformatics

PatBLAST

PatentInformatics

PatGenDB :

Automatically compiled DB of genomic sequences

DDBJ Databank of Japan

NCBI Natl Ctr for Biotechnology Information

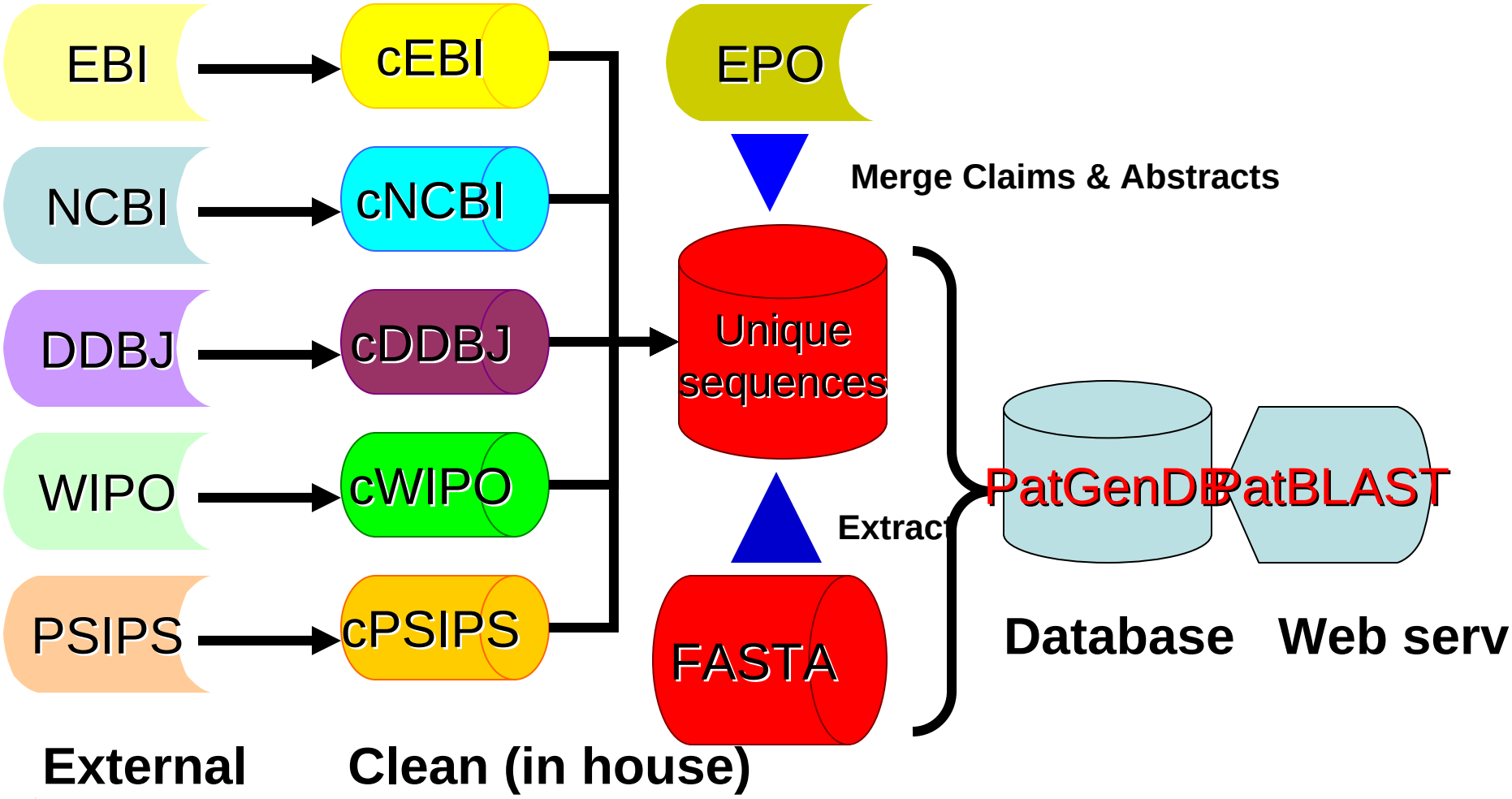
EBI European Bioinformatics Insitute

PSIPS USPTO Long sequence listings

WIPO PCT electronic sequence filings

Non redundant sequences or patents from public DBs

Steps to Create PatGenDB



Automated Compilation

Public DBs contain errors and inconsistencies

Can create significant problems for their search

Manual Curation

Time and resources consuming

Artificial Intelligence Curation

Avoid human error

Faster, cheaper to implement

Less complete than manual curation

Standards Facilitate Compilation

WIPO Standard ST.25 – Appendix 1 (WIPO/ PSIPS)

Mandatory and optional fields

Fields marked with numeric identifiers

100s General Patent Information

applicant, title, etc.

200s General Sequence Information

length, type, organism

300s Publication Information

journal, accession number

400 Sequence

Errors Cause Filing Problems

Errors in filing can invalidate an application

Common sources of error when filing:

- **Use of MS Word instead of plain text**
- **Legal aids make final revisions (should be scientists)**
- **Generate sequence listings by machine**
- **Cut and paste in listings, etc.**

R. Jones Nat. Biotech. 21:1239 (2003)

We are dealing with other side of the problem. Data may not comply with Standards and result in incomplete searches.

Data should be cleaned and curated

Misspelling

```
<110><81>@RIKEN
<120>  Method of analyses of protein-protein interactions
<130>
<160>  1
<210>  1
<211>  424
<212>  PRT
<213>  E.coli
<400>  1
Trp Phe Gly Asn Met Asn Val Leu Thr Phe leu Arg Asp Ile Gly Lys
1               5                   10                  15
His Phe Ser Val Asn Gln Met Ile Asn Lys Glu Ala Val Lys Gln Arg
            20                  25                  30
```

A protein element can be misread when automatically compiling a DB

WO06-057391

Corrupted Sequence

<400> 64
gtagtaaaac tcacggaagc gctaaaccac ttgatgag 38
Sequences from table 3:
AF030428 NM_006474
MWKVSALLFVLGSASLWVLAEGASTGQPEDDTETTGLEGGVAMP
GAEDDVVTPGTSEDRYKSGLTTLVATSVNSVTGIRIEDLPTSESTVHAQEQSPSATAS
NVATSHSTEKVDGDTQTTVEKDGLSTVTLVGIIVGVLLAIGFIGGIIVVVMRKMSGRY
SP

```
     1 gataaatgct gactccgctc ggaaagttct caactgcaaa gtttgctgtc cggctgccta
    61 gggtctggga agctcgggca ccctccctct ccggggctcc tgctcccacc cctccggccc
   121 ccccaccgtc gcgctcctcc aggctgggcc tgtggccgcg gtgcttttta attttccccc
   181 agctcagaat cttgctgctc ggcccccagg agagcaacaa ctcaacggga acgatgtgga
   241 aggtgtcagc tctgctcttc gttttgggaa gcgcgtcgct ctgggtcctg gcagaaggag
   301 ccagcacagg ccagccagaa gatgacactg agactacagg tttggaaggc ggcgttgcca
   361 tgccaggtgc cgaagatgat gtggtgactc caggaaccag cgaagaccgc tataagtctg
   421 gcttgacaac tctggtggca acaagtgtca acagtgtaac aggcattcgc atcgaggatc
   481 tgccaacttc agaaagcaca gtccacgcgc aagaacaaag tccaagcgcc acagcctcaa
   541 acgtggccac cagtcactcc acggagaaag tggatggaga cacacagaca acagttgaga
   601 aagatggttt gtcaacagtg accctggttg gaatcatagt tggggtctta ctagccatcg
   661 gcttcattgg tggaatcatc gttgtggtta tgcgaaaaat gtcgggaagg tactcgccct
   721 aaagagctga agggttacgc cctgctgcca acgtgcttaa aaaaagaccg tttctgactc
```

WO03-080640

Extraneous Characters

```
gtcgagcggt agaggcggat tttgtttttg agggcggcgg cggcggcggc ggcggcggcg     180
gcggtagggt cggggttttt cgggtcgggg t   539
<210> 129
<21gtttagg     240
gcggagttgt acgggcgagg tagtaagtgg ggcgtcgttg gcgggcgcgg cggtcgttgt     300
tatggattgg tgatcgcggc ggtttttgtt tcgttttttt cggtcgggcg tttgtttttt     360
agcgttaagt ttttagtcgg ggttatgggt tcgtcggcgg tcgcggcggc ggcggcggcg     420
```

Typographical Errors

<130>
<150> JP 2003-346248
<151> 2003-10-3
<150> JP 2004-212255
<151> 2004-7-20
<160> 16
<210> 1
<211> 20
<212> DNA
<213> Artificial sequence
<220>

WO05-033298

Corrupted Block Number letter L instead of Numeral 1

Nucleotide and Aminoacids mixed

```
caatcgcggg aagccagggt ttccagctag gacacagcag gtcgtgatcc gggtcgggac   15300
actgcctggc agaggctgcg agc atg ggg ccc tgg ggc tgg aaa ttg cgc       15350
                          met gly pro trp gly trp lys leu arg
                          -21 -20                  -15
tgg acc gtc gcc ttg ctc ctc gcc gcg gcg ggg act gca g gtaaggcttg    15400
trp thr val ala leu leu leu ala ala ala gly thr ala v
        -10                 -5               -1   1
```

WO06-067254

Page Numbers Misplaced

```
tgg acc gtc gcc ttg ctc ctc gcc gcg gcg ggg act gca g gtaaggcttg 15400
trp thr val ala leu leu leu ala ala ala gly thr ala v
        -10                 -5                -1   1
                                    8
ctccaggcgc cagaataggt tgagagggag cccccggggg gcccttggga atttattttt   15460
ttgggtacaa ataatcactc catccctggg agacttgtgg ggtaatggca cggggtcctt   15520
cccaaacggc tggaggggggc gctggagggg ggcgctgagg ggagcgcgag ggtcgggagg   15580
```

WO06-067254

Line Numbers in front of Blocks

```
SEQUENCE LISTING
     <110>   Applied Research Systems ARS Holding N.V.
 5   <120>   SPLICE VARIANT OF UNC5H2
     <130>   WO855
     <150>   EP04102511.5
10   <151>   2004-06-04
     <160>   11
     <170>   PatentIn version 3.3
15   <210>   1
```

WO05-118641

Altoris

PatentInformatic

Repeated Versions

<210> 45

<211> 9

<212> PRT

<213> Artificial Sequence

<223> Description of Artificial Sequence: hTERT-derived Synthetic Peptide

<400> 45

Arg Phe Ile Pro Lys Pro Asp Gly Leu
1 5

<210> 45

<211> 9

<212> PRT

<213> Artificial Sequence

<223> Description of Artificial Sequence: hTERT-derived Synthetic Peptide

<400> 46

Asp Phe Leu Leu Val Thr Pro His Leu
1 5

Unreadable Characters

<210> 6

<211> 60

<212> DNA

<213> Rattus norvegicus

<400> 6

actttcacaa gagctaagaa agctgcacag gtgaccatcc gttcttcggg
cacattttct 60

<210> <82>V

<211> 60

<212> DNA

<213> Rattus norvegicus

<400> <82>V

tgtgtccccc cgatgacttg gctgagcgag gactcttgga tatcgagact
tgcttctatg 60

<210> 8

Protein Sequence Labeled DNA

<212> DNA
<213> Eschierichia coli W3110
<400> 124

```
Met Gly Gln Glu Lys Leu Tyr Ile Glu Lys Glu Leu Ser Trp Leu Ser
  1               5                  10                  15
Phe Asn Glu Arg Val Leu Gln Glu Ala Ala Asp Lys Ser Asn Pro Leu
             20                  25                  30
Ile Glu Arg Met Arg Phe Leu Gly Ile Tyr Ser Asn Asn Leu Asp Glu
         35                  40                  45
Phe Tyr Lys Val Arg Phe Ala Glu Leu Lys Arg Arg Ile Ile Ile Ser
     50                  55                  60
```

WO06-001382

Unexpected Information

<210> SEQ ID NO 1

<211> LENGTH: N/A

<212> TYPE: N/A

<213> ORGANISM: N/A

<400> SEQUENCE: 1

The wrong sequence listing was accidentally published for this document. The application is being republished and a new sequence listing is coming soon.

US20020110548A1

Mixed Formats

Upper part in Alternative PSIPS format

```
(2) INFORMATION FOR SEQ ID NO: 4:
     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 27 amino acids
          (B) TYPE: amino acid
          (C) STRANDEDNESS: single
```

Bottom Part Regular PSIPS WIPO Format

```
<120> TITLE OF INVENTION: GENES INVOLVED IN INTESTINAL
INFLAMMATORY DISEASES AND USE THEREOF
<130> FILE REFERENCE: 37991-0009
<140> CURRENT APPLICATION NUMBER: US/10/240,046A
<141> CURRENT FILING DATE:  2003-04-02
<150> PRIOR APPLICATION NUMBER: PCT/FR 01/00935
<151> PRIOR FILING DATE: 2001-03-27
```

US20030190704A1

Automatically Compiled Database

Automated Procedure resulted in:

> 97,000 Patents

> 24.7 M unique sequences

Number of Sequences and sequence per patent contrib

EMBL	**2.8 M**	**30**	
NCBI	**1.9 M**	**25**	
DDBJ	**1.9 M**	**25**	
WIPO	**4.9 M**	**2777**	
PSIPS	**21.5 M**		**7979**

Database Uniqueness

Overlap of sequences in pairs of Dbs

	EMBL	DDBJ	NCBI	WIPO	PSIPS
EMBL	100	0	0	89	95
DDBJ	31	100	1	84	96
NCBI	30	0	100	92	96
WIPO	82	80	80	100	85
PSIPS	63	61	61	33	100

82% of the EMBL sequences is not found in WIPO
89% of the WIPO sequences is not found in EMBL

Database Uniqueness

Number of sequences found solely in that database

EMBL	**496085**	**18%**
NCBI	**0**	**0%**
DDBJ	**0**	**0%**
WIPO	**1470890**	**30%**
PSIPS	**17602116**	**82%**

Pharmacologically Relevant

Only a small number of patents claim current pharmacological targets:

Kinase	**6%**
GPCRs	**0.5%**
Ion Channel	**0.4%**
NHR	**0.07%**
Protease	**5%**
Reductase	**2%**

Summary

Public Databases contain errors

- Compromise applications
- Compromise searches

Automated compilation and cleaning of DBs

- Possible, efficient (not as as thorough manual)
- Artificial Intelligence can aid in the process

Databases show redundancies but all contain unique sequences

- At a given time
- EBI can substitute for NCBI and DDBJ

Bioinformatics research can benefit from DBs

- Sequences are related to pharma relevant targets
- Patents can aid in systems biology research

PatentInformatics

An Altoris, Inc. Project

Hugo O. Villar, Ph.D.

hugo@altoris.com

www.PatentInformatics.com

www.altoris.com