Boston 2007 PIUG Biotechnology

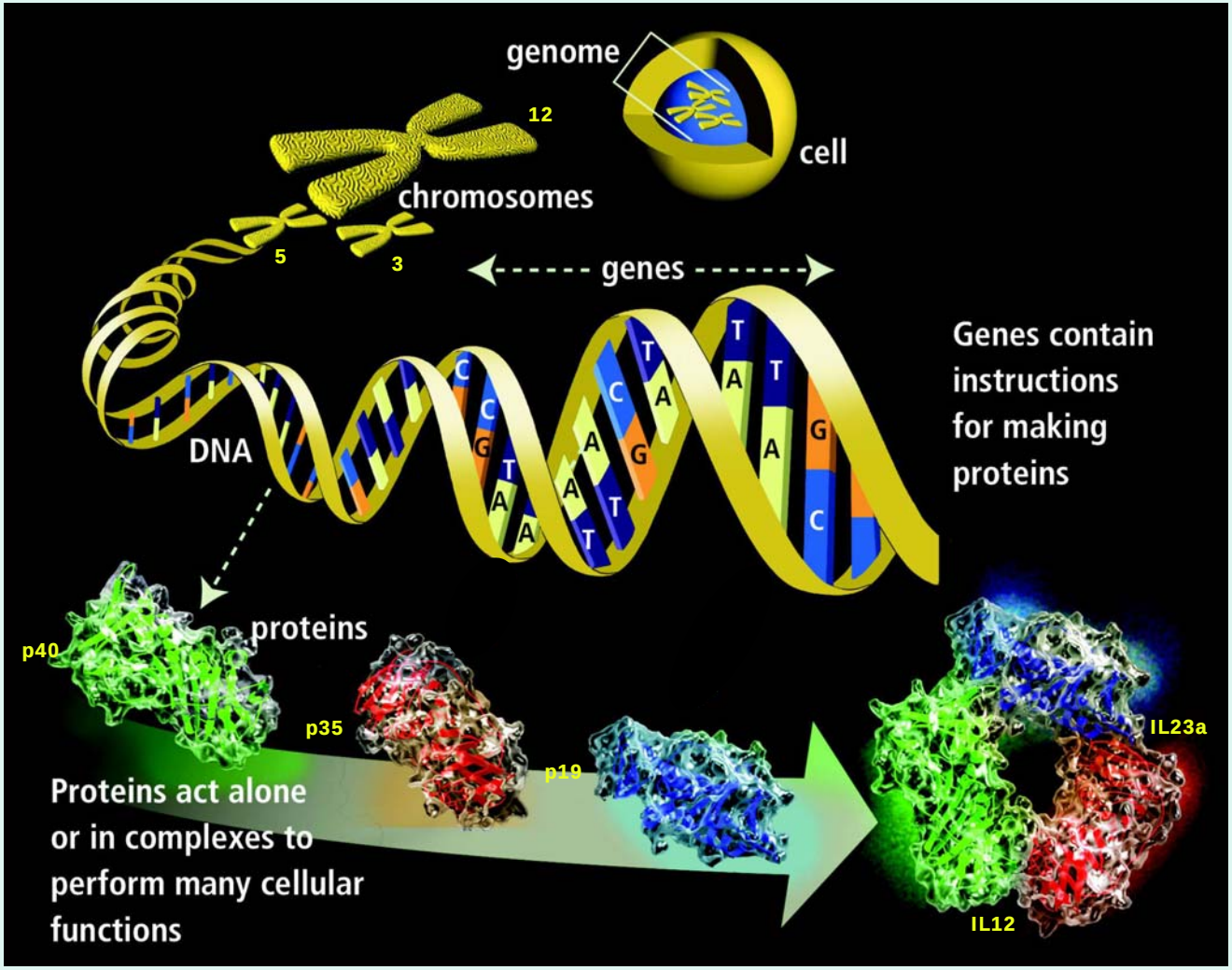
Genomics-Based Intellectual Property Protection

Understanding and Editing of Genomic Repetitive Sequences within Nucleic Acid Strings

Hypothetical question:

Please perform a global Freedom To Operate *(Patents)* search to find disclosures of *Anti-*IL12 Antibody *Mab*1 indicating cancer tumor defeat?

Genome



DOE www.ornl.gov/hgmis

Approach

- DATABASES
  - DERWENT GeneSeq
  - CAS Registry
  - PCTgen
  - *EMBL/NCBI Patent Division*
  - *PSIPS*
  - USGENE
  - *Other special databases?*
- ALGORITHMS
  - Local Alignment
    - ° BLAST (Basic Local Alignment Tool)
    - ° FASTA
    - ° BESTFIT

Approach *(cont.)*

- ALGORITHMS
  - Global Alignment
    - ° GAP
    - ° CLUSTAL W
  - Identity (Edit metric)
    - ° GenePAST

Sequence-based

- SEQUENCES

  - **IL12** protein

    For each subunit- p35, p40

  - **IL12** mRNA/DNA

    For each subunit- p35, p40

    Interested in family members?
    Interested in comparative genomes?
    Particular variants or polymorphisms?
    Functional domains?
    Promoter region?

  - ***Mab1*** heavy chain protein
  - ***Mab1*** light chain protein
  - ***Mab1*** heavy chain DNA
  - ***Mab1*** light chain DNA

    Variable and/or Fc-constant regions?
    Any special Fc substitutions or designs?

Pre-Search Manipulations

- Protein and DNA sequences first run within a text editor e.g. ReadSeq for size summation and reformatting

  http://iubio.bio.indiana.edu/soft/molbio/readseq/java/

- **IL12** mRNA/DNA?
  - Annotated within the RepeatMasker algorithm for genomic interspersed repetitive nucleic acid strings

    http://woody.embl-heidelberg.de/repeatmask/

    http://www.repeatmasker.org/

  - *Why?* The search sequences can then be edited for removal of confounding genomic interspersed repeats to avoid false positive records from entering sets

RepeatMasker

• Algorithm screens DNA for interspersed repeats and low complexity DNA sequences

• Output gives detailed annotation and shows modified version of the query sequence in which all the annotated repeats have been masked

• On average, nearly 50% of a human genomic DNA sequence is masked by the program

• Human genome is $2.9X10^9$bp

- What are the genomic repeats?

Low-complexity DNA

• Simple repeats (micro-satellites) can originate at any site in the genome, and therefore have an interspersed character

• Other low-complexity DNA, primarily poly-purine/poly-pyrimidine stretches, or regions of extremely high AT or GC content will result in spurious matches in BLAST

• Only all di- to tetrameric and some pentameric simple repeats are scanned for

LINEs

• LINE-1 (L1) Long Interspersed Nucleotide Elements are highly abundant in mammalian genomes

• Estimated 100,000 copies in the human genome

• Complete and transpositionally active L1 element is 6000 to 7000bp long

• Have 5'-untranslated region (UTR), two open reading frames, one of which (ORF2) has been shown to code for a reverse transcriptase, and a 3'-UTR followed by an A-rich tail

LINEs *(cont.)*

• Most L1 copies are 5'-truncated, often to the point that they comprise only part of the 3'-UTR

• New L1 copies originate in the genome *via* reverse transcription of RNA transcripts from active elements called source or master genes

• Few source genes have been active (in parallel) at any one time during evolution, giving rise to distinct subfamilies of pseudogenes as substitutions take place in the source gene sequence

SINEs

• Short Interspersed Nucleotide Elements (SINEs) are highly abundant in mammalian genomes

• The term SINE has come to be restricted to short retroposons with internal RNA polymerase III promoter sites in a region derived from a structural 7SL RNA pseudogene

• Our haploid genome contains more than a million of the 100- to 300-bp short interspersed repetitive SINE elements

MIRs(MERs)

- Called MIR for mammalian-wide interspersed repeat or MER for medium reiterated frequency sequence

- MIRs are classic, tRNA-derived SINEs

- High divergence and their presence at orthologous sites in different mammals indicate that MIRs, at least in part, amplified before the mammalian radiation

- Estimated 300,000 copies discernible, which account for 1 to 2% of our DNA

Alus

- The Alu family of short interspersed repeated DNA elements (SINEs) are distributed throughout primate genomes

- Amplified via an RNA-mediated transposition process to a copy number of about 500,000

- Comprising an estimated 5% of the human genome

- Derive their name from a single recognition site for the endonuclease Alu 1 located near the middle of the Alu sequence

LTRs

- Retrovirus-like elements, retroposons characterized by long terminal repeats

- Carry the transcriptional regulatory sequences and are reproduced partly from each other in a complex reverse transcription process

- Hundreds-of-thousands of LTRs are found in the human genome

- Endogenous retroviral LTR sequences range from 300-1000bp

MaLRs

- Named derived from ‘Mammalian apparent LTR- retrotransposons’
- MaLRs are 2000-3000bp elements in an abundant superfamily of mammalian LTR- transposons
- Estimated 40,000 to 100,000 members in primate genomes
- MaLRs were distributed before the radiation of eutherian mammals 80-100 million years ago

Combined Search Strategy

- Sequence-based discovery and analysis
  - Protein level
  - DNA level ***(edited)***
- Conceptual (keyword/ classification)

International Patent Searching

VALIDITY INFRINGEMENT PATENTABILITY PATENT WATCHES

TECHNOLOGY & PATENT RESEARCH
International℠

Bringing the power of all these resources to your search

Hand Searching of Full Collections from:

- United States
- EPO European Patent Office
- Patent Cooperation Treaty
- Japan
- United Kingdom
- Germany
- France
- The Netherlands
- Switzerland
- Sweden
- Canada
- Australia
- Austria
- Belgium
- Denmark
- Italy
- Luxembourg
- Russia
- Other Countries

International Databases, Patents and Non-Patents:

- Derwent World Patents Index
- IFI CLAIMS Databases
- US, Intl., EPO Classifications
- Full Text Patent Databases
- INPADOC
- Chemical Abstracts, Registry
- Chemical Structures/Markush
- Chemical Reactions, Beilstein
- Biotechnology, Biosequences
- Medline, Biomedical Files
- Scientific & Patent Citations
- Scientific Trade Journals
- Dissertations, Conferences
- IEEE/INSPEC
- Electrical Engineering Jnls.
- IBM, Bell Labs. Disclosures
- Business News and Journals

Worldwide Libraries and Resources:

- Patent Office Files from US, UK, EPO, JP, etc.
- USPTO Libraries
- EPO Libraries, The Hague
- Japanese Libraries/Institutes
- British Science Ref. Library
- Natl. Institutes of Health, NIH
- Natl. Library of Medicine NLM
- Library of Congress
- Elec./Engineering Institutes
- University Libraries Worldwide
- University Theses Collections
- The Internet, and more…

In Many Languages:

- Japanese, German, French, Italian, Spanish, Russian, etc.

TPR – Searching for the Professionals, by the Professionals™

- **Searches and Patent Watches – flexible to your requirements and budget**
- **In-depth investigation of alternative sources of prior art for top-priority licensing, validity and/or litigation cases**

17055 Via Del Campo Suite 200 ▲ San Diego CA 92127
tel: 858.592.9084 ▲ fax: 858.592.0045 ▲ searches@TPRInternational.com

Steve Chappell Mitchell
Technology and Patent Research International
searches@tprinternational.com
858-592-9084

end