Searching for antibody information on STN®

Robert Austin – FIZ Karlsruhe
Agenda

• Introduction to antibodies and immunoglobulins
• Understanding and searching antibody indexing in DGENE, REGISTRY and CAplus<sup>SM</sup>
• Sequence searching for antibodies using DGENE, USGENE<sup>®</sup>, PCTGEN and REGISTRY
  – Complementarity Determining Regions (CDRs) using BLAST<sup>®</sup> and Sequence Code Match (SCM) searching
  – Multi-file search and post-processing

See also: Sequence Basics e-Seminar (June 2010):
http://www.stn-international.com/Sequence_Basics_Seminar.html
Why use STN to search for antibodies?

• Sequence information on STN is comprehensive
  – Four sequence databases allow users to achieve a comprehensive search of sequences from both patent and non-patent documents

• Sequence databases on STN are timely
  – Allows you to keep up-to-date with the most current information

• Sequence indexing is unique
  – Allows you to retrieve sequences containing uncommon residues or chemical modifications that are difficult to find
Four sequence databases on STN provide unique information

• **DGENE**  
  – Sequences from the 41 authorities covered by DWPI\textsuperscript{SM}  
  – Sequence data are intellectually analyzed and indexed  
  – Legal status and patent family display options  
  – Information is updated once every two weeks

• **REGISTRY**  
  – Sequences from the 61 authorities covered by CAplus  
  – Sequences also come from >3000 life science journals  
  – Sequence data are intellectually analyzed and indexed  
  – Information is updated daily
Four sequence databases on STN provide unique information (cont.)

- **USGENE**
  - Sequences from all relevant USPTO published patent applications and granted (issued) patents
  - Legal status and patent family display options
  - Updated weekly, within three days of publication

- **PCTGEN**
  - Sequences submitted and published electronically as a formal part of WIPO/PCT published patent applications
  - Legal status and patent family display options
  - Updated weekly, within 24 hours of publication
Antibodies are produced as a defence against foreign substances (antigens)

- Antibodies (Ab) are specialised glycoproteins, which differ in size, charge, carbohydrate content and amino acid sequence composition
- They are also known as immunoglobulins (Ig)
- Antibodies are found in blood and other bodily fluids of mammals and some other vertebrates
- They are a central part of the humoral immune response (HIR) and are synthesised by B-cells
Antibodies are useful because of their biological properties and high specificity

• There are different classes of antibodies, depending on their structure
  – Mammals have five classes of antibodies
    • α (IgA), δ (IgD), ε (IgE), γ (IgG), μ (IgM)
  – Each class has different biological properties

• Antibodies are highly specific to antigens
  – Able to locate one molecule of a protein antigen out of more than $10^8$ similar molecules
  – Useful in targeted therapy and as diagnostic tools
Mammalian antibodies are Y-shaped and composed of heavy and light chains

- Antibodies are composed of four polypeptides
  - Two identical light chain (L)
  - Two identical heavy chain (H)
- Both light and heavy chains consist of constant (C) region domains with little variability, and variable (V) region domains with high variability
- The four chains are held together by several disulphide bonds and form a Y-shaped molecule
- The antigen binding sites (CDRs) are located at the tips of the Y-shaped arms
Additional nomenclature describes variable region domains

• Light chains exist in two forms
  – kappa (κ), lambda (λ)

• Heavy chains exist in five forms
  – α (IgA), δ (IgD), ε (IgE), γ (IgG), μ (IgM)
  – Variation in heavy chains gives rise to various antibody subclasses: IgG1, IgG2, IgA1, etc.
Mammalian antibodies are Y-shaped and composed of heavy and light chains.

- **Antigen binding site**
- **Light Chain (L)**: κ, λ
- **Heavy Chain (H)**: γ, μ, α, δ, ε

**Antigen binding site**

**Hinge**

**CDR** = complementarity determining region

**V_L** = variable region - light chain

**V_H** = variable region - heavy chain

**C_L** = constant region - light chain

**C_H** = constant region - heavy chain
Humanization of antibodies is an important process for therapeutic usage

• Immunotherapy (or biotherapy)
  – Uses certain parts of the immune system to fight diseases such as cancer
  – Treatments are less toxic and potentially more effective than chemical drugs
  – Types of antibodies used for therapy
    • Monoclonal antibodies
    • Chimeric antibodies
    • CDR-Grafted antibodies
    • Phage Display antibodies
Antibodies are also used as diagnostic tools

• Antibody tools reduce assay time without compromising sensitivity
  – Flow cytometric analysis
    • Analysis of morphological complexity of the cells, DNA content (cell cycle analysis), cell sorting
  – Microarray technology
    • Proteomics: Protein characterization and analysis of diseased vs. healthy patients
  – Immunoblotting (or western blot)
    • Detection of a specific protein from a tissue or cell sample
  – Immunohistochemistry
    • Localization of protein(s) in cells or tissue sections using antibodies
Agenda

• Introduction to antibodies and immunoglobulins

• Understanding and searching antibody indexing in DGENE, REGISTRY, and CAPI

• Sequence searching for antibodies using DGENE, USGENE®, PCTGEN and REGISTRY
  – Complementarity Determining Regions (CDRs) using BLAST® and Sequence Code Match (SCM) searching
  – Multi-file search and post-processing
Antibody sequences are indexed in GENESEQ™ on STN (DGENE)

- Description (/DESC)
  - Concise one-line description of the sequence
  - E.g. Mouse anti-protein X antibody VL region

- Keyword (/KW) indexing for antibody sequences
  - **Type**, e.g. humanized, monoclonal
  - **Region**, e.g. light chain constant region
  - **Activity**, e.g. antibody therapy, immune stimulation
  - **Target**, e.g. protein X
  - **Disease**, e.g. immune disorder, autoimmune disease
  - **Technology**, e.g. antibody engineering, antibody array

- Abstract (AB)
  - Includes the use of the antibody within the invention

- Features Table (/FEAT)
  - Details about Domain, Region, Disulphide-bonds, etc
Each DGENE record has a concise one-line description of the antibody sequence.

<table>
<thead>
<tr>
<th>L1</th>
<th>ANSWER 1 OF 1</th>
<th>DGENE COPYRIGHT 2010 THOMSON REUTERS on STN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCESSION NUMBER:</td>
<td>AEN02775 protein</td>
<td>DGENE</td>
</tr>
<tr>
<td>TITLE:</td>
<td>New antibody useful for treating e.g. cancer competitively inhibiting binding of competitor antibody having complementarity determining region of specific amino acid sequences as given in the specification to G-protein coupled receptor.</td>
<td></td>
</tr>
<tr>
<td>INVENTOR:</td>
<td>Howard M; Schall T</td>
<td></td>
</tr>
<tr>
<td>PATENT ASSIGNEE:</td>
<td>(CHEM-N)CHEMOCENTRX INC.</td>
<td></td>
</tr>
<tr>
<td>PATENT INFO:</td>
<td>WO 2006116319 A2 20061102 60</td>
<td></td>
</tr>
<tr>
<td>APPLICATION INFO:</td>
<td>WO 2006-US15492 20060419</td>
<td></td>
</tr>
<tr>
<td>PRIORITY INFO:</td>
<td>US 2005-674140P 20050421</td>
<td></td>
</tr>
<tr>
<td>PAT. SEQ. LOC:</td>
<td>Disclosure; SEQ ID NO 22</td>
<td></td>
</tr>
<tr>
<td>DATA ENTRY DATE:</td>
<td>22 FEB 2007 (first entry)</td>
<td></td>
</tr>
<tr>
<td>DOCUMENT TYPE:</td>
<td>Patent</td>
<td></td>
</tr>
<tr>
<td>LANGUAGE:</td>
<td>English</td>
<td></td>
</tr>
<tr>
<td>DESCRIPTION:</td>
<td>anti-CCX-CK2-antibody 11G8 VL region SEQ ID NO 22.</td>
<td></td>
</tr>
</tbody>
</table>
Each DGENE record has keyword indexing for the antibody sequence

| KEYWORD: | cytostatic; neuroprotective; nootropic; nephrotropic; antirheumatic; antiarthritic; cardiant; antiarteriosclerotic; antiasthmatic; dermatological; antiinflammatory; gastrointestinal-gen.; antipsoriatic; vasotrophic; immunosuppressive; antiulcer; ophthalmological; antidiabetic; vulnerary; hepatotropic; anorectic; respiratory-gen.; gynecological; hemostatic; cardiovascular-gen.; contraceptive; protein interaction; antibody; angiogenesis inhibition; cell proliferation; protein detection; antibody therapy; arthritis; Alzheimers disease; multiple sclerosis; renal failure; rheumatoid arthritis; transplant rejection; asthma; glomerulonephritis; contact dermatitis; inflammatory bowel disease; colitis; psoriasis; reperfusion injury; ocular disease; diabetic retinopathy; retinopathy of prematurity; macular degeneration; graft rejection; neovascular glaucoma; rubeosis; Osier-Webber Syndrome; telangiectasis; angiofibroma; Crohns disease; eczema; wound healing; osteopathic; fractures; burns; inflammation; ischemia; peripheral vascular disease; pre-eclampsia; cardiovascular disease; 11G8; light chain variable region. |
| ORGANISM: | Mus sp. |
Each DGENE abstract describes the use of the antibody sequence within the invention

**ABSTRACT:**

The invention describes an antibody (A1) that competitively inhibits binding of a competitor antibody (a1) to CCX-CKR2 (G-protein coupled receptor), where the competitor antibody comprises the complementarity determining region (CDR) of specific amino acid sequences as given in the specification. The antibody is useful in a pharmaceutical composition for inhibiting angiogenesis or proliferation of a cancer cell in an individual such as other than human having or pre-disposed to have arthritis; for detecting a cell expressing CCX-CKR2 in a biological sample; for treating Alzheimer's disease; multiple sclerosis; kidney dysfunction; rheumatoid arthritis; cardiac allograft rejection; atherosclerosis; asthma; glomerulonephritis; contact dermatitis; inflammatory bowel disease; colitis; psoriasis; reperfusion injury; ocular angiogenic diseases, for example, • • • joints (e.g. arthritis and hemophiliac joints), healing of wounds, fractures, and burns, inflammatory diseases, ischemic heart, and peripheral vascular diseases; preclampsia and cardiovascular disease; for birth control. The antibody competitively inhibits binding of a competitor antibody to CCX-CKR2, and potently inhibits angiogenesis. This is the amino acid sequence of anti-CCX-CK2-antibody 11G8 light chain variable region.
The DGENE Feature Table includes detailed annotations for the antibody sequence

<p>| AMINO ACID COUNTS: 2 A; 4 R; 2 N; 6 D; 0 B; 2 C; 5 Q; 3 E; 0 Z; 9 G; 3 H; 5 I; 10 L; 4 K; 1 M; 4 F; 6 P; 15 S; 5 T; 1 W; 6 Y; 7 V; 0 Others |
| SEQUENCE LENGTH: 100 |
| SEQUENCE |
| 1 dvlmtqtpls lpvslgdqas iscrsshryv hsdgntylew ylkpgqspk |
| 51 lliykvsnrf sgvpdrfsgs gsgtdftlki srveaedlg yycfgqgshvp |
| FEATURE TABLE: |</p>
<table>
<thead>
<tr>
<th>Key</th>
<th>Location</th>
<th>Qualifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>1..23</td>
<td>note</td>
</tr>
<tr>
<td>Region</td>
<td>24..39</td>
<td>note</td>
</tr>
<tr>
<td>Region</td>
<td>40..54</td>
<td>note</td>
</tr>
<tr>
<td>Region</td>
<td>55..61</td>
<td>note</td>
</tr>
<tr>
<td>Region</td>
<td>62..93</td>
<td>note</td>
</tr>
</tbody>
</table>
Antibody light and heavy chain sequences are indexed in separate DGENE records

| L1 | ANSWER 1 OF 2  DGENE COPYRIGHT 2010 THOMSON REUTERS on STN |
| AN | AWO22746 protein DGENE |
| TI | Purifying an antibody from a composition comprises loading the... |
| DESC | Anti-VEGF bevacizumab humanized antibody light chain sequence, SEQ ID 12. |
| KW | protein purification; VEGF ligand; cation-exchange; chromatography; light chain; humanized antibody; protein purification; vascular endothelial growth factor. |
| SQL | 214 |

| L1 | ANSWER 2 OF 2 DGENE COPYRIGHT 2010 THOMSON REUTERS on STN |
| AN | AWO22745 protein DGENE |
| TI | Purifying an antibody from a composition comprises loading the... |
| DESC | Anti-VEGF bevacizumab humanized antibody heavy chain sequence, SEQ ID 11. |
| KW | protein purification; VEGF ligand; cation-exchange; chromatography; heavy chain; humanized antibody; protein purification; vascular endothelial growth factor. |
| SQL | 453 |

Note: This example comes from WO2009058812.

Thomson Reuters indexing makes clear which one is which.
Antibody light and heavy chain sequences are indexed in separate USGENE records.

<table>
<thead>
<tr>
<th>=&gt;</th>
<th>D AN TRIAL 1-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2</td>
<td>ANSWER 1 OF 2  USGENE COPYRIGHT 2010</td>
</tr>
<tr>
<td>AN</td>
<td>20090148435.12  Protein  USGENE</td>
</tr>
<tr>
<td>TI</td>
<td>ANTIBODY PURIFICATION BY CATION EXCHANGE CHROMATOGRAPHY  (PublishedApplication)</td>
</tr>
<tr>
<td>DESC</td>
<td>Artificial Protein; <strong>Sequence is synthesized</strong>; sequence 12 of 20</td>
</tr>
<tr>
<td>MTY</td>
<td>Protein</td>
</tr>
<tr>
<td>SQL</td>
<td>214</td>
</tr>
</tbody>
</table>

| L2 | ANSWER 2 OF 2  USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN |
| AN | 20090148435.11 |
| TI | ANTIBODY PURIFICATION BY CATION EXCHANGE CHROMATOGRAPHY  (PublishedApplication) |
| DESC | Artificial Protein; **Sequence is synthesized**; sequence 11 of 20 |
| MTY | Protein |
| SQL | 453 |

**Note:** This example comes from US20090148435 A1, which is equivalent to WO2009058812 A1.

Patent applicants often do not provide a clear description.
Antibodies are indexed as substances in CAS REGISTRY

• Antibodies are indexed as sequences if the sequence(s) is provided by the author(s)
  – The Note (/NTE) field contains additional information about the sequence (i.e. chemically modified, linkages, uncommon amino acids, etc.)
  – Separate records for the full multi-chain antibody, light chain and heavy chain sequences may be created

• Antibody sequences are also indexed with
  – Index Names
  – Trade names
  – Generic names
  – Lab names
CA Index Names provide information, such as sequence type, organisms, and strain/cell/tissue types.

The CA Index Name for Avastin contains additional information, such as the isotype (G1), the antigen (VEGF), monoclonal antibody, etc.

Brand/generic names, and lab names are listed under “Other names”.

Avastin is registered as a multichain sequence (two heavy and two light chains).
Modifications and/or linkages between chains are listed in the /NTE field.

<table>
<thead>
<tr>
<th>type</th>
<th>location</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bridge</td>
<td>Cys-22</td>
<td>Cys-96</td>
</tr>
<tr>
<td>bridge</td>
<td>Cys-150</td>
<td>Cys-206</td>
</tr>
<tr>
<td>bridge</td>
<td>Cys-226</td>
<td>Cys-214''</td>
</tr>
<tr>
<td>bridge</td>
<td>Cys-232</td>
<td>Cys-232'</td>
</tr>
<tr>
<td>bridge</td>
<td>Cys-235</td>
<td>Cys-235'</td>
</tr>
<tr>
<td>bridge</td>
<td>Cys-267</td>
<td>Cys-327</td>
</tr>
<tr>
<td>bridge</td>
<td>Cys-373</td>
<td>Cys-431</td>
</tr>
<tr>
<td>bridge</td>
<td>Cys-22'</td>
<td>Cys-96'</td>
</tr>
<tr>
<td>bridge</td>
<td>Cys-150'</td>
<td>Cys-206'</td>
</tr>
<tr>
<td>bridge</td>
<td>Cys-226'</td>
<td>Cys-214''</td>
</tr>
<tr>
<td>bridge</td>
<td>Cys-267'</td>
<td>Cys-327'</td>
</tr>
<tr>
<td>bridge</td>
<td>Cys-373'</td>
<td>Cys-431'</td>
</tr>
<tr>
<td>bridge</td>
<td>Cys-23''</td>
<td>Cys-88''</td>
</tr>
<tr>
<td>bridge</td>
<td>Cys-134''</td>
<td>Cys-194''</td>
</tr>
</tbody>
</table>

In REGISTRY, the specific residue(s) position(s) are listed in the /NTE field.
CAS will index both light and heavy chains for antibodies

<table>
<thead>
<tr>
<th>SEQ</th>
<th>1</th>
<th>EVQLVESGGG</th>
<th>LVQPGGSLRL</th>
<th>SCAASGYTFT</th>
<th>NYGMNWVRQA</th>
<th>PGKGLEWVGW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>51</td>
<td>INTYGEPTY</td>
<td>AADFKRRFTF</td>
<td>SLDTSKSTAY</td>
<td>LQMNSLRAED</td>
<td>TAVYYCAKYP</td>
</tr>
<tr>
<td></td>
<td>101</td>
<td>HYYGSSHWYF</td>
<td>DVWGQGTLVT</td>
<td>VSSASTKGPS</td>
<td>VFPLAPSSKS</td>
<td>TSGGTAALGC</td>
</tr>
<tr>
<td></td>
<td>401</td>
<td>PPVLDSDGSF</td>
<td>FLYSKLTVDK</td>
<td>SRWQQGNVFS</td>
<td>CSVMEHALHN</td>
<td>HYTQKSLSLS</td>
</tr>
<tr>
<td></td>
<td>451</td>
<td>PGK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEQ</td>
<td>1</td>
<td>DIQMTQSPSS</td>
<td>LSASVGDRVT</td>
<td>ITCSASQDIS</td>
<td>NYLNWYQQKP</td>
<td>GKPVKVLIVF</td>
</tr>
<tr>
<td></td>
<td>51</td>
<td>TSSLHSGVPS</td>
<td>RFSGSGSGTD</td>
<td>FTLTISSLQP</td>
<td>EDFATYYCQQ</td>
<td>YSTVPWTFGQ</td>
</tr>
<tr>
<td></td>
<td>201</td>
<td>LSSPVTKSFN</td>
<td>RGEC</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| SEQ  | 1  | DIQMTQSPSS| LSASVGDRVT | ITCSASQDIS | NYLNWYQQKP | GKPVKVLIVF |
|      | 51 | TSSLHSGVPS| RFSGSGSGTD | FTLTISSLQP | EDFATYYCQQ | YSTVPWTFGQ |
|      | 201| LSSPVTKSFN| RGEC       |            |            |            |
Light and heavy chain sequences may also be indexed in separate REGISTRY records.

Tip: Compare to DGENE (slide 19) and to USGENE (slide 20).

Note: REGISTRY sequence records from patents, often do not include a description of the sequence.
The antibody REGISTRY numbers are indexed in CAplus bibliographic records.

<table>
<thead>
<tr>
<th>PATENT NO.</th>
<th>KIND</th>
<th>DATE</th>
<th>APPLICATION NO.</th>
<th>DATE</th>
</tr>
</thead>
</table>

A method for purifying an antibody by cation exchange chromatography is described in which a high pH wash step is used to remove of contaminants prior to eluting the desired antibody using an elution buffer with increased conductivity. Preferably the antibody binds human CD20, such as rituximab, or binds human vascular endothelial growth factor (VEGF), such as bevacizumab...
The antibody Registry Numbers are linked to detailed roles and index terms in CAplus

CC  15-3 (Immunochemistry)
Section cross-reference(s): 9

IT  174722-31-7P, Rituximab  216974-75-3P, Bevacizumab
RL: ARG (Analytical reagent use); BPN (Biosynthetic preparation); PUR (Purification or recovery); THU (Therapeutic use); ANST (Analytical study); BIOL (Biological study); PREP (Preparation); USES (Uses)
  (antibody purification by cation exchange chromatog. using high pH wash step to remove of contaminants prior to eluting in buffer with increased conductivity)

IT  192433-87-7  214551-08-3  214551-09-4  214551-11-8  214551-12-9
214551-13-0  444104-00-1  556112-97-1  556112-98-2  556112-99-3
556113-00-9  1150529-46-6  1150802-74-6  1150802-75-7  1150802-76-8
1150802-77-9  1150802-78-0  1150802-79-1  1150802-80-4  1150802-81-5
RL: PRP (Properties)
  (unclaimed protein sequence; antibody purification by cation exchange chromatog. using a high pH wash step to remove of contaminants prior to eluting in buffer with increased conductivity)

This CAplus record has indexing for both the multi-chain antibody, and separately for the light and heavy chain sequences.
Antibodies are indexed in CAplus

• Covered by controlled terms in CAplus
  – Consult the Lexicon for old and new terms
  – Refine with additional concepts contained within the same index term by using the (L) operator

• CAS Registry Numbers (CAS RNs) are available to supplement a CAplus search, if retrieval of specific antibody substances is required
How are antibodies indexed in CAplus?

• Antibodies are indexed to the most specific level disclosed in a document
  – Light chains
    • κ (kappa), λ (lambda)
  – Heavy chains
    • α (IgA), δ (IgD), ε (IgE), γ (IgG), μ (IgM)
    • Subclasses: IgG1, IgG2, IgA1, etc.

• Descriptors (limiters) can provide additional information
  – Examples: bispecific, catalytic, labeled, neutralizing, humanized, chimeric, etc.
Antibody controlled term indexing has changed over time in CAplus

<table>
<thead>
<tr>
<th>Controlled Term</th>
<th>Years of Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antibodies and Immunoglobulins/CT</td>
<td>2002 to present</td>
</tr>
<tr>
<td>Amboceptors/CT</td>
<td>1907-1946</td>
</tr>
<tr>
<td>Antibodies/CT</td>
<td>1907-2001</td>
</tr>
<tr>
<td>Globulins, immune/CT</td>
<td>1967-1976</td>
</tr>
<tr>
<td>Immunoglobulins/CT</td>
<td>1962-2001</td>
</tr>
</tbody>
</table>
Immunoglobulin (Ig) nomenclature can be used to focus on specific forms of interest.

**Search Question:**
Find records covering lambda light chain immunoglobulins. Are any of these IgG immunoglobulins?

Identify uses of these substances.
Search Strategy

To find references on immunoglobulins...

Step 1. Search appropriate Ig terms
Step 2. Refine with class, subclass, or chain nomenclature
Step 3. Evaluate using D SCAN HIT
Step 4. (Optional) Refine with additional concepts or CAS Roles
Step 5. Display results
Tips for searching in CAplus

• Use BOTH single word immunoglobulin and antibody terms in the /IT field, especially to include records prior to 2002
• Use (L) proximity to add modifying terms
• Specific subclass may be used but allow for generic class for comprehensive results
  – For example, IgG2a is of most interest but IgG2 encompasses it and should be included
• Supplement CAS Roles with text terms in the modifier
Use SET commands to automatically add plurals and abbreviations

=> SET PLU ON; SET ABB ON; SET SPE ON

  SET COMMAND COMPLETED

=> S (IMMUNOGLOBULIN(L) LAMBDA)/IT

  11999 IMMUNOGLOBULIN/IT
  143185 IMMUNOGLOBULINS/IT
  150880 IMMUNOGLOBULIN/IT
  ((IMMUNOGLOBULIN OR IMMUNOGLOBULINS)/IT)
  20421 IG/IT
  6316 IGS/IT
  25278 IG/IT
  ((IG OR IGS)/IT)
  156119 IMMUNOGLOBULIN/IT
  ((IMMUNOGLOBULIN OR IG)/IT)
  178328 LAMBDA
  68 LAMBDAS
  178342 LAMBDA
  (LAMBDA OR LAMBDAS)
  L1  1213 IMMUNOGLOBULIN/IT (L) LAMBDA
Adding antibody term retrieves additional answers

\[
\Rightarrow S \text{ ANTIBODY/IT (L) LAMBDA}
\]

\[
70328 \text{ ANTIBODY/IT }
\]

\[
214281 \text{ ANTIBODIES/IT }
\]

\[
223724 \text{ ANTIBODY/IT }
\]

\[
((\text{ANTIBODY OR ANTIBODIES})/\text{IT})
\]

\[
178328 \text{ LAMBDA }
\]

\[
68 \text{ LAMBDAS}
\]

\[
178342 \text{ LAMBDA}
\]

\[
((\text{LAMBDA OR LAMBDAS})
\]

\[
L2 \quad 568 \text{ ANTIBODY/IT (L) LAMBDA}
\]

\[
\Rightarrow S \text{ L1 OR L2}
\]

\[
L3 \quad 1374 \text{ L1 OR L2}
\]
Alternative search using **BOTH** terms

=> S (ANTIBODY OR IMMUNOGLOBULIN)/IT (L) LAMBDA
L4  1374 (ANTIBODY OR IMMUNOGLOBULIN)/IT (L) LAMBDA

=> S L4 (L) IGG?
  80343 IGG?
L5  58 L4 (L) IGG?

=> D HIT SCAN

L5  58 ANSWERS  CAPLUS COPYRIGHT 2010 ACS on STN
IT  Antibodies and Immunoglobulins
   RL: ADV (Adverse effect, including toxicity); BSU (Biological study, unclassified); DGN (Diagnostic use); BIOL (Biological study); USES (Uses)
   (IgG, .lambda., .kappa.; gammopathy detected by serum protein electrophoresis for predicting and managing therapy of lymphoproliferative disorder in liver transplant recipients)

HOW MANY MORE ANSWERS DO YOU WISH TO SCAN? (1): **END**
Remove the kappa light chain entries from the index term search

=> S L5 (NOTL) KAPPA
    70588 KAPPA
    11 KAPPAS
    70594 KAPPA
    (KAPPA OR KAPPAS)
L6    32 L5 (NOTL) KAPPA

=> D SCAN HIT

L6    32 ANSWERS

IT  Antibodies and Immunoglobulins
    RL: BSU (Biological study, unclassified); BIOL (Biological study)
        (light chain, .lambda.; of IgG autoantibodies to TSH receptors in Graves disease of humans)

HOW MANY MORE ANSWERS DO YOU WISH TO SCAN? (1): 2
Index terms describe therapeutic applications

Lymphoma
(B-cell, monoclonal antibodies to IgG1
\(\lambda\) lambda. paraprotein variable region epitopes in diagnosis of human)

Antibodies
RL: BIOL (Biological study)
(monoclonal, IgG1.\(\lambda\). paraprotein variable region epitopes recognition by, of human, lymphoid disease diagnosis in relation to)

Immunoglobulins
RL: PRP (Properties)
(mapping of \(\lambda\). light chain epitopes for human lupus IgG autoantibodies)

Protein sequences
(of Ig \(\lambda\). light chains of humans in relation to lupus IgG autoantibody binding)

HOW MANY MORE ANSWERS DO YOU WISH TO SCAN? (1): 0
Agenda

• Introduction to antibodies and immunoglobulins
• Understanding and searching antibody indexing in DGENE, REGISTRY, and CAplusSM
• Sequence searching for antibodies using DGENE, USGENE ®, PCTGEN and REGISTRY
  – Complementarity Determining Regions (CDRs) using BLAST® and Sequence Code Match (SCM) searching
  – Multi-file search and post-processing
Sequence searching for CDRs

• Within the variable domains of the heavy and light chains, there are three hyperactivity regions called Complementarity Determining Region (CDR)
• Each chain contains three CDR regions
• CDR is the region that recognizes the different antigens
  – CDR is important when developing antibodies
BLAST and/or Sequence Code Match (SCM) can be used to retrieve CDRs

Search Question:
The epithelial cell adhesion molecule (EpCAM) is a cell surface protein that is expressed by a variety of tumor cells. We have identified CDR sequence \textcolor{ForestGreen}{DMGWGS\textcolor{Orange}{GWRP} YYYYGMDV} in our laboratory. Find all patent publications that disclose this sequence or similar sequences.
Search Strategy for DGENE, USGENE, PCTGEN and REGISTRY/CAplus

Step 1. RUN BLAST in USGENE, DGENE and PCTGEN using offline BATCH mode

Step 2. Repeat the search using CAS REGISTRY BLAST

Step 3. Retrieve, merge, organize by patent family and display USGENE, DGENE and PCTGEN results

Step 4. Retrieve, identify and display unique CAplus references from the REGISTRY BLAST search

Step 5. Post-process results into tables and reports
RUN BLAST searches in DGENE, USGENE and PCTGEN in offline BATCH mode

Tip: DGENE, USGENE and PCTGEN BLAST searches can be run in parallel using BATCH mode.
Adjust the REGISTRY BLAST settings for optimal search retrieval

**Recommended BLAST settings for short sequences (35 amino acids or less)**
1) Turn OFF (uncheck) the low complexity filter
2) Increase Expectation Value to 20,000
3) Decrease Word Size to 2
4) Choose the PAM-30 Weight Matrix

Increase the maximum number of answers to 1,000.

For more information about BLAST matrices, visit the NCBI web site.
Retrieve references for sequences

**Note**: in this example, BLAST sequences with scores of 42 or more (60% match or more) are selected.
Retrieve REGISTRY BLAST results

Transfer BLAST sequences with scores of 42 or more (60% match or more).

=> FIL REGISTRY
=> QUE (1065745-06-3 OR 1067695-29-7 OR 1065745-11-0 OR . . .
L1  QUE (1065745-06-3 OR 1067695-29-7 OR 1065745-11-0 OR . . .
=> QUE (862861-57-2 OR 487483-96-5 OR 215027-98-8 OR . . .
L2  QUE (862861-57-2 OR 487483-96-5 OR 215027-98-8 OR . . .
=> QUE (1089240-08-3 OR 1074003-68-1 OR 960549-36-4 OR . . .
L3  QUE (1089240-08-3 OR 1074003-68-1 OR 960549-36-4 OR . . .
=> S L1 OR L2 OR L3
L4  36 L1 OR L2 OR L3

Commands within the dotted box are automatic commands.

36 similar sequences (L4) with BLAST scores of 42 or more.
Tip: BLAST is better than SCM for searching short sequences for less than 100% match

=> FIL REG

=> S DMGWGSGWPRYYYYGMDV/SQSFP
 L5    20 DMGWGSGWPRYYYYGMDV/SQSFP

=> S L5 NOT L4
 L6    0 L5 NOT L4

=> DEL L5-L6 Y

Subsequence family protein search (/SQSFP) (L5), allows for amino acid family substitution.

REGISTRY BLAST (L4) retrieved 16 extra sequences with 60% or higher match by score that were not retrieved with /SQSFP (L5).
Retrieve the DGENE, USGENE and PCTGEN in BLAST search results

=> **FILE DGENE;** RUN GETBATCH EPCAMCDR

.....

ENTER (ALL) OR ? : 60%

L5  RUN STATEMENT CREATED
L5  32 DMGWGSGWRPYYYYGMDV/SQP.-F F -E 20000 -W 2 -M PAM30

=> **FILE USGENE;** RUN GETBATCH EPCAMCDR

.....

ENTER (ALL) OR ? : 60%

L6  RUN STATEMENT CREATED
L6  17 DMGWGSGWRPYYYYGMDV/SQP.-F F -E 20000 -W 2 -M PAM30

=> **FILE PCTGEN;** RUN GETBATCH EPCAMCDR

.....

ENTER (ALL) OR ? : 60%

L7  RUN STATEMENT CREATED
L7  12 DMGWGSGWRPYYYYGMDV/SQP.-F F -E 20000 -W 2 -M PAM30

DGENE, USGENE and PCTGEN BLAST searches are retrieved with the RUN GETBATCH command.
Merge and review the DGENE, USGENE and PCTGEN in BLAST search results

=> DUP IDE L5 L6 L7

FILE 'DGENE' ENTERED AT . . .
COPYRIGHT (C) 2010 THOMSON REUTERS

FILE 'USGENE' ENTERED AT . . .
COPYRIGHT (C) 2010 SEQUENCEBASE CORP

FILE 'PCTGEN' ENTERED AT . . .
COPYRIGHT (C) 2010 WIPO

L8 61 DUP IDE L5 L6 L7 (INCLUDES 0 SETS OF DUPLICATES)
ANSWERS '1-32' FROM FILE DGENE
ANSWERS '33-49' FROM FILE USGENE
ANSWERS '50-61' FROM FILE PCTGEN

=> SOR SCORE D AN D
PROCESSING COMPLETED FOR L8
L9 61 SOR L8 SCORE D AN D

The multi-file answers (L8) can be into descending BLAST score order (L9).

STN
Review the BLAST search results

Example: displaying the best answer from each database in a free-of-charge review format.
Review the BLAST search results (cont.)

L9

ANSWER 21 OF 61

USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN

TI Method of identifying binding site domains that retain the capacity of binding to an epitope (Patent)

DESC Homo Sapiens Protein; sequence 54 of 77

MTY Protein

SQL 127

SCORE 70 100% of query self score 70

BLASTALIGN

Query = 18 letters
Length = 127
Score = 69.8 bits (157), Expect = 2e-18
Identities = 18/18 (100%), Positives = 18/18 (100%)

Query: 1   DMGWGSGWRPYYYYGMDV 18
   DMGWGSGWRPYYYYGMDV

Sbjct: 99  DMGWGSGWRPYYYYGMDV 116

L9

ANSWER 25 OF 61

PCTGEN COPYRIGHT 2010 WIPO on STN

TI CROSS-SPECIES-SPECIFIC BINDING DOMAIN

MTY PRT

SQL 504

SCORE 70 100% of query self score 70

BLASTALIGN

Query = 18 letters
Length = 504
Score = 69.8 bits (157), Expect = 5e-18
Identities = 18/18 (100%), Positives = 18/18 (100%)

Query: 1   DMGWGSGWRPYYYYGMDV 18
   DMGWGSGWRPYYYYGMDV

Sbjct: 99  DMGWGSGWRPYYYYGMDV 116
Use the STN Express 8.4 Patent Family Manager wizard display the results

Access the patent family manager wizard from the Discover! Menu.

Choose a bibliographic display format with alignment for the first (best) hit, and a free-of-charge format with alignment for the rest of the sequences in each patent family group.
The patent family manager begins by organising the results using FSORT...

```plaintext
=> FSORT L9

L10 61 FSO L9

11 Multi-record Families

<table>
<thead>
<tr>
<th>Family</th>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family 1</td>
<td>1-18</td>
</tr>
<tr>
<td>Family 2</td>
<td>19-20</td>
</tr>
<tr>
<td>Family 3</td>
<td>21-22</td>
</tr>
<tr>
<td>Family 4</td>
<td>23-25</td>
</tr>
<tr>
<td>Family 5</td>
<td>26-27</td>
</tr>
<tr>
<td>Family 6</td>
<td>28-29</td>
</tr>
<tr>
<td>Family 7</td>
<td>30-34</td>
</tr>
<tr>
<td>Family 8</td>
<td>35-42</td>
</tr>
<tr>
<td>Family 9</td>
<td>43-56</td>
</tr>
<tr>
<td>Family 10</td>
<td>57-58</td>
</tr>
<tr>
<td>Family 11</td>
<td>59-60</td>
</tr>
</tbody>
</table>

1 Individual Record | Answer 61
0 Non-patent Records
```

FSORT organizes the patent sequence records by Publication, Application, Related, and Priority numbers.

In this example, 12 patent family groups (i.e. 11 + 1) are retrieved.

Commands in **RED** are those issued automatically by the STN Express Patent Family Manager.
...and then continues by displaying the family groups in the specified formats

<table>
<thead>
<tr>
<th>=&gt; DIS L10 PFAM=4 1 BIB,PSL,SQL,SCORE,ALIGN</th>
</tr>
</thead>
<tbody>
<tr>
<td>L10</td>
</tr>
<tr>
<td>AN</td>
</tr>
<tr>
<td>TI</td>
</tr>
<tr>
<td>IN</td>
</tr>
<tr>
<td>PA</td>
</tr>
<tr>
<td>PI</td>
</tr>
<tr>
<td>AI</td>
</tr>
<tr>
<td>PRAI</td>
</tr>
<tr>
<td>LA</td>
</tr>
<tr>
<td>OS</td>
</tr>
<tr>
<td>DESC</td>
</tr>
<tr>
<td>PSL</td>
</tr>
<tr>
<td>SQL</td>
</tr>
<tr>
<td>SCORE</td>
</tr>
<tr>
<td>BLASTALIGN</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Query: 1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Sbjct: 99</td>
</tr>
</tbody>
</table>

Commands in RED are those issued automatically by the STN Express Patent Family Manager.
...and then continues by displaying selected results in the specified formats (cont.)

=> DIS L10 PFAM=4 2-TOT TRIAL,SCORE,ALIGN

L10  ANSWER 24 OF 61 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STNFAMILY4
TI    Anti-EpCam Immunoglobulins (PublishedApplication)
DESC  Artificial Protein; Anti-EpCAM Heavy Chain; sequence 1 of 2
MTY   Protein
SQL   457
SCORE 70 100% of query self score 70
BLASTALIGN
  Query  = 18 letters
  Length = 457
  Score  = 69.8 bits (157), Expect = 5e-18
  Identities = 18/18 (100%), Positives = 18/18 (100%)
  Query: 1  DMGWGSGWRPYYYYGMDV 18
           DMGWGSGWRPYYYYGMDV
  Sbjct:  99 DMGWGSGWRPYYYYGMDV 116

L10  ANSWER 25 OF 61 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STNFAMILY4
TI    Anti-EpCAM immunoglobulins (PublishedApplication)
MTY   Protein
SQL   457
SCORE 70 100% of query self score 70
BLASTALIGN
  Query  = 18 letters
  Length = 457
  Score  = 69.8 bits (157), Expect = 5e-18
  Identities = 18/18 (100%), Positives = 18/18 (100%)
  Query: 1  DMGWGSGWRPYYYYGMDV 18
           DMGWGSGWRPYYYYGMDV
  Sbjct:  99 DMGWGSGWRPYYYYGMDV 116

These two USGENE hits are in the same family as the DGENE record on the previous slide.
...and then continues by displaying selected results in the specified formats (cont.)

This USGENE record is the “individual record” in the FSORT answer set (L10).
Retrieve and identify any unique CAplus references by the REGISTRY search

=> FILE HCAPPLUS

=> TRA L10 PN

L11 TRANSFER L10 1- PN : 22 TERMS
L12 16 L11
ALL TERMS IN L11 RETRIEVED.

=> S L4
L13 14 L4

=> S L13 NOT L12
L14 1 L13 NOT L12

Transfer Publication Numbers (PN) from DGENE, USGENE and PCTGEN (L10) to find corresponding HCAplus records (L12).

The 36 REGISTRY records (L4) correspond to 14 HCAplus records (L13).

In this example, one additional relevant reference has been found by including a REGISTRY/CAplus search (L14).
Display any unique CAplus references retrieved in the REGISTRY search

=> D L14 BIB ABS HITRN

L14 ANSWER 1 OF 1 HCAPLUS COPYRIGHT 2010 ACS on STN
AN 2008:1210508 HCAPLUS Full-text
DN 149:446284
TI Cross-species bispecific single-chain antibodies to human or non-chimpanzee primate CD3e and surface antigen for treating tumorous, proliferative, or immunological disease and cancer
IN Ebert, Evelyn; Meier, Petra; Sriskandarajah, Mirnaalini; Burghart, Elke; Wissing, Sandra; Klinger, Matthias; Bluemel, Claudia; Raum, Tobias; Rau, Doris; Mangold, Susanne; Kvesic, Majk; Kischel, . . . . Hausmann, Susanne; Riethmueller, Gert
PA Micromet A.-G., Germany
CODEN: PIXXD2
DT Patent
LA English
FAN.CNT 3

<table>
<thead>
<tr>
<th>PATENT NO.</th>
<th>KIND</th>
<th>DATE</th>
<th>APPLICATION NO.</th>
<th>DATE</th>
</tr>
</thead>
</table>

. . . .
The present invention provides polypeptides comprising an antibody binding domain capable of binding to an epitope of human and non-chimpanzee primate CD3 e-chain fused to a cell surface antigen selected from epidermal growth factor receptor, epidermal growth factor receptor variant III, melanoma chondroitin sulfate proteoglycan, carbonic anhydrase IX, CD30, CD33, CD44 variant 6, EpCAM, Her2/neu, MUC1, and IgE. An N-terminal 1-27 amino acid residues polypeptide fragment of the extracellular domain of human CD e-chain was identified which — in contrast to all other known epitopes of CD3 e — maintains its 3-dimensional structural integrity when taken out of its native environment in the CD3 complex. . . . . Further, the invention provides a method for the identification of polypeptides comprising a cross-species specific binding domain capable of binding to an epitope of human and non-chimpanzee primate CD3 e-chain. It is important to include the HITRN or HITIND display for post-processing.

(amino acid sequence; cross-species bispecific single-chain antibodies to human or non-chimpanzee primate CD3e and surface antigen for treating tumorous, proliferative, or immunol. disease and cancer)
STN Express post-processing tools provide the finishing touches to the multi-file search.

1) DGENE, USGENE and PCTGEN results \((L10)\) can be conveniently tabulated using the STN Express Table Tool and exported to a Microsoft Excel worksheet.

2) The REGISTRY “BLAST Report with Alignment data” tool merges BLAST alignments with corresponding unique CAplus records \((L14)\) to form a single RTF file.
1) DGENE, USGENE and PCTGEN results can be tabulated and exported to Excel.

Preferred fields, fonts, labels, etc, can be saved as a Template for repeated re-use.
Once in Excel there are various options to sort, filter and review the multi-file results.

Some tips for Microsoft Excel:
- Resize columns and rows as desired – especially the BLAST alignment column to approx 77
- View, Freeze panes – holds the top row fixed when scrolling down
- Add Filters – provides a great way to navigate results – for example by BLAST score (above)
2) REGISTRY BLAST Alignments can be merged with corresponding CAplus records

Preferred fields, fonts, labels, etc., can be saved as a Template for repeated re-use.
The REGISTRY “BLAST Report with Alignment data” tool provides an RTF file.
# Patent classifications for antibody topics

<table>
<thead>
<tr>
<th>USCL Class</th>
<th>Subclass</th>
<th>IPC8 Subclass</th>
<th>Group</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>530</td>
<td>387.1+</td>
<td>C07K</td>
<td>16/00+</td>
<td>Immunoglobulins, Antibodies</td>
</tr>
<tr>
<td>530</td>
<td>388.1+</td>
<td>C12P</td>
<td>21/08+</td>
<td>Monoclonal Antibodies</td>
</tr>
<tr>
<td>530</td>
<td>391.1+</td>
<td>C07K</td>
<td>16/00+</td>
<td>Conjugated antibodies</td>
</tr>
<tr>
<td>424</td>
<td>1.49</td>
<td>A61K</td>
<td>49/00+</td>
<td>Compositions containing radio-labeled antibodies</td>
</tr>
<tr>
<td>424</td>
<td>130.1</td>
<td>A61K</td>
<td>39/395+</td>
<td>Compositions for body treatment containing antibodies (therapeutics, vaccines)</td>
</tr>
<tr>
<td>435</td>
<td>7.1+</td>
<td>G01N</td>
<td>33/53+</td>
<td>Immunoassays using antibodies</td>
</tr>
<tr>
<td>435</td>
<td>188</td>
<td>C12N</td>
<td>09/96+</td>
<td>Antibodies conjugated to enzymes</td>
</tr>
<tr>
<td>435</td>
<td>188.5</td>
<td>C07K</td>
<td>16/00+</td>
<td>Catalytic antibodies</td>
</tr>
<tr>
<td>435</td>
<td>325+</td>
<td>C12N</td>
<td>5/00+</td>
<td>Cells that express antibodies (fused, recombinant)</td>
</tr>
<tr>
<td>525</td>
<td>54.1</td>
<td>A61K</td>
<td>47/48+</td>
<td>Antibodies bound to resins</td>
</tr>
</tbody>
</table>
Derwent World Patents Index® (DWPI\textsuperscript{SM})
Manual Codes (/MC) for antibody topics

More than 120 Manual Codes (/MC) are available for antibody searching.

http://scientific.thomson.com/cgi-bin/mc/search.cgi
Summary

- DGENE provides detailed annotations and indexing for text searching for antibody technologies
- REGISTRY provides extensive annotations, and common, trade, generic, and lab antibody names
- CAS controlled and index terms are all useful for retrieving antibody information in CAplus
  - Use text terms to search for types of antibodies in CAplus
    - Class (α (IgA), δ (IgD), ε (IgE), γ (IgG), μ (IgM))
- BLAST and SCM searches are available for antibody sequence searching in DGENE, USGENE, PCTGEN and REGISTRY on STN
Resources

• Archived e-Seminars:
  www.cas.org/support/stngen/stntraining/recorded.html
  – Unmasking the World of Antibodies in CAS REGISTRY
  – Finding Antibodies and Immunoglobulins
  – Sequence motif searching on STN

• STN User Documentation:
  www.cas.org/support/stngen/stndoc/sequences.html
  – Quick Reference Guides
    • CAS REGISTRY: BLAST similarity searching via STN Express
    • CAS REGISTRY: Exact and pattern searching of nucleic acid sequences
    • CAS REGISTRY: Exact and pattern searching of protein sequences

• Sequence Searching on STN public workshop
  www.stn-international.com/sequence_searching.html
Searching for antibody information on STN®

www.stn-international.com