Integrating Sequence Results with Multiple Patent Sources to Provide Better Reports and Visualizations

Monica Weiss-Nolen, MSIS, Sanofi Pasteur
John Willmore, BizInt Solutions

PIUG Biotechnology Conference February 2015

Requested Enhancements

Issue	Requests
Company Names	• Filter out Inventors • Standardize names
Patent Family	• Simplify legal status per country • Visualize status per company
Integrating Sequence Data	• How to combine multiple sequence records into a single row effectively?

www.bizint.com/slides

Working with Company Names

- Issues:
  Inconsistent format
  Inventor names
  Subsidiary units
  Corporate ownership
  OCR errors
  Transliteration

Patent Assignee	Database
CREW B CELL HOLLAND BUOY CRUCELL US DEPARTMENT OF HEALTH & HUMAN SERVICES US GOVERNMENT US NAVY	FAMPAT
CRUCELL HOLLAND B.V. ; SULLIVAN, NANCY ; CHAKRABARTI, BIMAL ; YANG, ZHI-YONG ; PAU, MARIA GRAZIA ; GOUDSMIT, JAAP ; NABEL, GARY J. ; THE GOVERNMENT OF THE UNITED STATES OF AMERICA, AS REPRESENTED BY THE SE CRETARY, DEPARTMENT OF HEALTH AND HUMAN SERVICES	GQPAT Gold+ Proteins
CRUCELL HOLLAND BV. US DEPT HEALTH & HUMAN SERVICES.	Derwent GeneSeq
National Institutes of Health Crucell Holland BV	Thomson Reuters Cortellis Patents

GLAKSOSMITKLAJN BAJOLODZHIKALZ GLAXOSMITHKLINE BIOLOGICALS

Working with Company Names

These issues can be fixed in all records using VP-SCE tools:

- List comparison
- Fuzzy match
- Thesaurus construction

Patent Assignee	Inventor(s)
CRUCELL HOLLAND BV	BIMAL CHAKRABARTI
PAU MARIA GRAZIA	GRAZIA PAU MARIA
GOUDSMIT JAAP	JAAP GOUDSMIT
NABEL GARY J	ZHI YONG YANG
NABEL GARY	YANG ZHI YONG
SULLIVAN NANCY	SULLIVAN NANCY
YANG ZHI YONG	PAU MARIA GRAZIA
US GOV HEALTH AND HUMAN SERV	NABEL GARY J
US GOVERNMENT	NABEL GARY
US HEALTH	GOUDSMIT JAAP
US GOV	CHAKRABARTI BIMAL
CHAKRABARTI BIMAL	NANCY SULLIVAN

CRUCELL HOLLAND BV
US DEPT HEALTH AND HUMAN SERVICES

vantagepoint
Smart Charts Edition

Software for Business Intelligence
BizInt Smart Charts

Working with Company Names

- With some databases, this is a "solved problem"
- From the earlier example…

CRUCELL HOLLAND BV. US DEPT HEALTH & HUMAN SERVICES.	Derwent GeneSeq

- Or from Thomson Reuters Cortellis, companies with roles…

Companies	
Company	**Relationship**
Crucell Holland BV	Patent Assignee/Owner
US Government	Patent Assignee/Owner
National Institutes of Health	Patent Assignee/Owner

Working with Company Names

- **Do you need to clean up every variation of a company name?**
- **BizInt Smart Charts Reference Rows** lets you select a single representative value
- Records grouped by shared publication numbers "Common Patent Family"

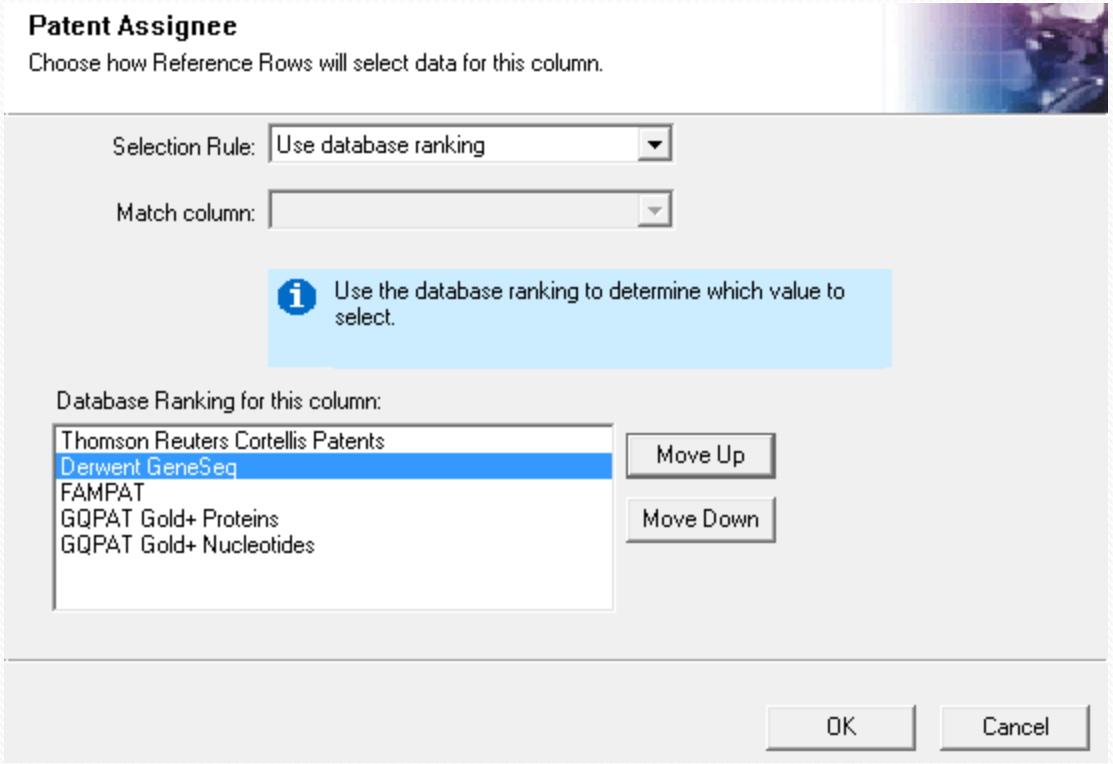
Database	Patent Assignee
FAMPAT	CREW B CELL HOLLAND BUOY CRUCELL US DEPARTMENT OF HEALTH & HUMAN SERVICES US GOVERNMENT US NAVY
Thomson Reuters Cortellis Patents	National Institutes of Health Crucell Holland BV
GQPAT Gold+ Proteins	
GQPAT Gold+ Proteins	CRUCELL HOLLAND B.V. ; SULLIVAN, NANCY ; CHAKRABARTI, BIMAL ; YANG, ZHI-YONG ; PAU, MARIA GRAZIA ; GOUDSMIT, JAAP ; NABEL, GARY J. ; THE GOVERNMENT OF THE UNITED STATES OF AMERICA, AS REPRESENTED BY THE SE CRETARY, DEPARTMENT OF HEALTH AND HUMAN SERVICES
GQPAT Gold+ Proteins	THE GOVERNMENT OF THE UNITED STATES OF AMERICA, ASREPRESE NTED BY THE SECRETARY, DEPARTMENT OF HEALTHAND HUMAN SERVICES (US) ; CRUCELL HOLLAND B.V. (NL)
GQPAT Gold+ Proteins	THE UNITED STATES OF AMERICAAS REPRESENTED BY THE DEPARTMENT OF HEALTH AND HUMAN SERVICES
GQPAT Gold+ Proteins	THE GOVERNMENT OF THE UNITED STATES OF AMERICAAS REPRESENTED BY THE SECRE TARY DEPARTMENT OF HEALTH AND HUMAN SERVICES
Derwent GeneSeq	CRUCELL HOLLAND BV. US DEPT HEALTH & HUMAN SERVICES.
Derwent GeneSeq	CRUCELL HOLLAND BV. US DEPT HEALTH & HUMAN SERVICES.
Derwent GeneSeq	CRUCELL HOLLAND BV. US DEPT HEALTH & HUMAN SERVICES.

Working with Company Names

- Select the value from one record to represent each family:
  - By database
  - By # entries
  - Manually

BizInt Smart Charts

Reference Rows™



Software for Business Intelligence
BizInt Smart Charts

Working with Company Names

CREW B CELL HOLLAND BUOY
CRUCELL
US DEPARTMENT OF HEALTH & HUMAN SERVICES
US GOVERNMENT
US NAVY
National Institutes of Health ✔
Crucell Holland BV

National Institutes of Health ✔
Crucell Holland BV

CRUCELL HOLLAND B.V. ; SULLIVAN, NANCY ; CHAKRABARTI, BIMAL ; YANG, ZHI-YONG ; PAU, MARIA GRAZIA ; GOUDSMIT, JAAP ; NABEL, GARY J. ; THE GOVERNMENT OF THE UNITED STATES OF AMERICA, AS REPRESENTED BY THE SE CRETARY, DEPARTMENT OF HEALTH AND HUMAN SERVICES
THE GOVERNMENT OF THE UNITED STATES OF AMERICA, ASREPRESE NTED BY THE SECRETARY, DEPARTMENT OF HEALTHAND HUMAN SERVICES (US) ; CRUCELL HOLLAND B.V. (NL)
THE UNITED STATES OF AMERICAAS REPRESENTED BY THE DEPARTMENT OF HEALTH AND HUMAN SERVICES
THE GOVERNMENT OF THE UNITED STATES OF AMERICAAS REPRESENTED BY THE SECRE TARY DEPARTMENT OF HEALTH AND HUMAN SERVICES
CRUCELL HOLLAND BV. US DEPT HEALTH & HUMAN SERVICES.
CRUCELL HOLLAND BV. US DEPT HEALTH & HUMAN SERVICES.
CRUCELL HOLLAND BV. US DEPT HEALTH & HUMAN SERVICES.

BizInt Smart Charts

Reference Rows™

Working with Company Names

- Selecting “cleaner” names before normalization reduces the work required
- Company name cleanup still required
  - Differences between database providers
  - Different entities filing with each authority/year
- Even the cleanest name doesn’t necessarily help
  - Crucell was acquired by Johnson & Johnson in 2010

Working with Legal Status

- TIP: let us know if something isn't right so we can fix it!

Database	Legal Status	Latest Legal Status	Legal Status
GQPAT Gold+ Proteins	Application	Public. of supplementary search report; 2012-04-11	
GQPAT Gold+ Nucleotides		Public. of supplementary search report; 2012-04-11	Application

Database	Pub. Status
GQPAT Gold+ Proteins	Application
GQPAT Gold+ Nucleotides	Application
FAMPAT	PENDING

Working with Legal Status

- FAMPAT has detailed legal status in ACT field

LEGAL DETAILS FOR WO2011071574: EED=2030-09-01; STATE=ALIVE; STATUS=PENDING
AD=2010-09-01 CO=WO/APP SI=Pos EG=EXM
Application details
APC=WOWOUS2010047586 APD=2010-09-01 XAP=2010WO-US47586
AD=2011-06-16 CO=WO/A2 SI=Pos EG=EXM
International application published with declaration under Article 17 (2) (a)
PC=WO PN=WO2011071574 KD=A2 PD=2011-06-16 XPN=WO201171574
AD=2011-10-06 CO=WO/A3 SI=Pos EG=EXM
Published application
PC=WO PN=WO2011071574 KD=A3 PD=2011-10-06 XPN=WO201171574
LEGAL DETAILS FOR DESIGNATED STATE EP2473525: EED=2014-08-27; STATE=DEAD; STATUS=LAPSED
Corresponding cc : CC=EPCorresponding appl: EP10836350 CAPD=2010-09-01 CAP=2010EP-0836350
Corresponding cc : CC=EPCorresponding pat: EP2473525 CKD=A2 CPD=2012-07-11 CPN=EP2473525
AD=2011-08-31 CO=WO/121 EG=DCS
EP: The EPO has been informed by wipo that ep was designated in this application
Corresponding cc: CC=EP
AD=2014-11-27 CO=EP/STCHG
Patent status changed by the national office
LEGAL DETAILS FOR DESIGNATED STATE US2012164153: EED=2030-09-01; STATE=ALIVE; STATUS=PENDING
Corresponding cc : CC=USCorresponding appl: US13393622 CAPD=2010-09-01 CAP=2010US-13393622
Corresponding cc : CC=USCorresponding pat: US2012164153 CKD=A1 CPD=2012-06-28

Working with Legal Status

- BizInt Smart Charts summarizes this as “Family Status”

Family Status			
Pub No.	State	Status	Expiry
WO2011071574	ALIVE	PENDING	2030-09-01
EP2473525	DEAD	LAPSED	2014-08-27
US2012164153	ALIVE	PENDING	2030-09-01

Working with Legal Status

- What is the current status of US 8101739?

Patent Family			Family Status			
Patent	Kind	Date	Pub No.	State	Status	Expiry
WO 200637038	A1	2006-04-06	WO2006037038	ALIVE	PENDING	2025-09-27
CA 2581840	A1	2006-04-06	AU2005289439	ALIVE	GRANTED	2025-09-27
AU 2005289439	A1	2006-04-06	CA2581840	ALIVE	GRANTED	2025-09-27
WO 200637038	A9	2006-05-26	EP1797113	ALIVE	GRANTED	2025-09-27
WO 200637038	B1	2006-08-03	IL182225	DEAD	LAPSED	2012-09-20
EP 1797113	A1	2007-06-20	IN2674/DELNP/2007	ALIVE	GRANTED	2025-09-27
IN 2007DN02674	A	2007-08-03	JP2008514203	ALIVE	GRANTED	2025-09-27
IL 182225	D0	2007-09-20	US2009232841	ALIVE	GRANTED	2027-06-07
JP 2008514203	A	2008-05-08	US2012156239	ALIVE	PENDING	2025-09-27
US 20090232841	A1	2009-09-17				
AU 2005289439	B2					
US 8101739	B2					
US 20120156239	A1					
JP 5046941	B2					
IN 259912	B					
CA 2581840	C	2014-08-05				
EP 1797113	B1	2014-11-26				

```
AD=2012-01-24 CO=US/B2 SI=Pos EG=PIF
Granted patent as second publication
PC=US PN=US8101739 KD=B2 PD=2012-01-24 XPN=US
AD=2012-01-24 CO=US/354 SI=Pos EG=PIF EG=SPC
```

Working with Legal Status

- **NEW!** Additional details now in Family Status

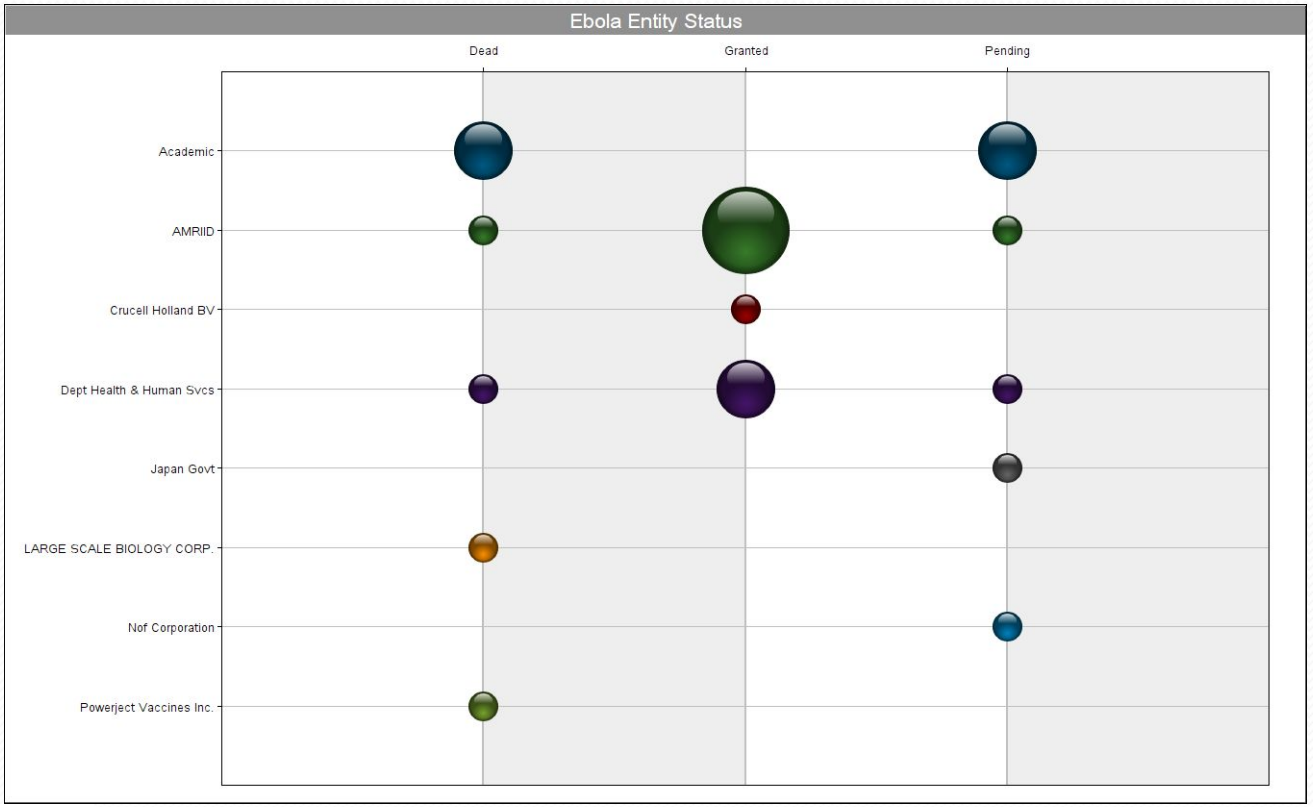
Patent Family			Family Status			
Patent	Kind	Date	Pub No.	State	Status	Expiry
WO 200637038	A1	2006-04-06	WO2006037038	ALIVE	PENDING	2025-09-27
CA 2581840	A1	2006-04-06	AU2005289439	ALIVE	GRANTED	2025-09-27
AU 2005289439	A1	2006-04-06	CA2581840	ALIVE	GRANTED	2025-09-27
WO 200637038	A9	2006-05-26	EP1797113	ALIVE	GRANTED	2025-09-27
WO 200637038	B1	2006-08-03	IL182225	DEAD	LAPSED	2012-09-20
EP 1797113	A1	2007-06-20	IN2674/DELNP/2007	ALIVE	GRANTED	2025-09-27
IN 2007DN02674	A	2007-08-03	JP2008514203	ALIVE	GRANTED	2025-09-27
IL 182225	D0	2007-09-20	US2009232841	ALIVE	GRANTED	2027-06-07
JP 2008514203	A	2008-05-08	US8101739	ALIVE	GRANTED	2027-06-07
US 20090232841	A1	2009-09-17	US2012156239	ALIVE	PENDING	2025-09-27
AU 2005289439	B2	2011-12-01				
US 8101739	B2	2012-01-24				
US 20120156239	A1	2012-06-21				
JP 5046941	B2	2012-10-10				
IN 259912	B	2014-04-04				
CA 2581840	C	2014-08-05				
EP 1797113	B1	2014-11-26				

Working with Legal Status

- Use VP-SCE to filter Patent Family and/or Family Status
- Select values for reporting or visualization
- Next challenge: annotating Patent Family with the status information from Family Status

Working with Legal Status

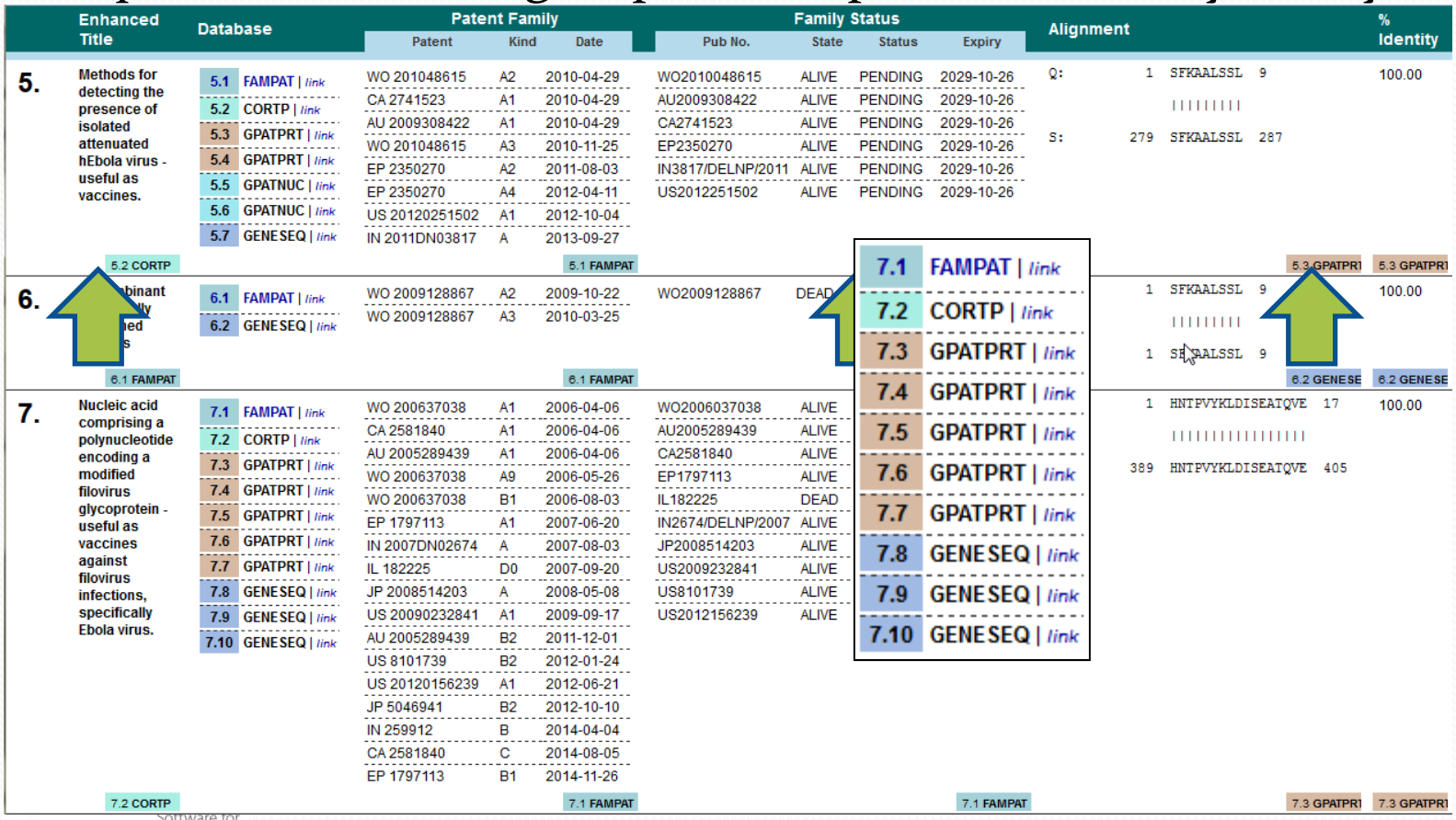
- Cleaned companies and status make visualizations possible



Software for Business Intelligence
BizInt Smart Charts

Working with Sequence Results

- Sequence results are grouped with patent results by family



	Enhanced Title	Database	Patent Family: Patent	Kind	Date	Family Status: Pub No.	State	Status	Expiry	Alignment	% Identity
5.	Methods for detecting the presence of isolated attenuated hEbola virus - useful as vaccines.	5.1 FAMPAT | link 5.2 CORTP | link 5.3 GPATPRT | link 5.4 GPATPRT | link 5.5 GPATNUC | link 5.6 GPATNUC | link 5.7 GENESEQ | link	WO 201048615 CA 2741523 AU 2009308422 WO 201048615 EP 2350270 EP 2350270 US 20120251502 IN 2011DN03817	A2 A1 A1 A3 A2 A4 A1 A	2010-04-29 2010-04-29 2010-04-29 2010-11-25 2011-08-03 2012-04-11 2012-10-04 2013-09-27	WO2010048615 AU2009308422 CA2741523 EP2350270 IN3817/DELNP/2011 US2012251502	ALIVE ALIVE ALIVE ALIVE ALIVE ALIVE	PENDING PENDING PENDING PENDING PENDING PENDING	2029-10-26 2029-10-26 2029-10-26 2029-10-26 2029-10-26 2029-10-26	Q: 1 SFKAALSSL 9 S: 279 SFKAALSSL 287	100.00
6.	...binant ...ily ...ed ...s	6.1 FAMPAT | link 6.2 GENESEQ | link	WO 2009128867 WO 2009128867	A2 A3	2009-10-22 2010-03-25	WO2009128867	DEAD			1 SFKAALSSL 9 1 SFKAALSSL 9	100.00
7.	Nucleic acid comprising a polynucleotide encoding a modified filovirus glycoprotein - useful as vaccines against filovirus infections, specifically Ebola virus.	7.1 FAMPAT | link 7.2 CORTP | link 7.3 GPATPRT | link 7.4 GPATPRT | link 7.5 GPATPRT | link 7.6 GPATPRT | link 7.7 GPATPRT | link 7.8 GENESEQ | link 7.9 GENESEQ | link 7.10 GENESEQ | link	WO 200637038 CA 2581840 AU 2005289439 WO 200637038 WO 200637038 EP 1797113 IN 2007DN02674 IL 182225 JP 2008514203 US 20090232841 AU 2005289439 US 8101739 US 20120156239 JP 5046941 IN 259912 CA 2581840 EP 1797113	A1 A1 A1 A9 B1 A1 A D0 A A1 B2 B2 A1 B2 B C B1	2006-04-06 2006-04-06 2006-04-06 2006-05-26 2006-08-03 2007-06-20 2007-08-03 2007-09-20 2008-05-08 2009-09-17 2011-12-01 2012-01-24 2012-06-21 2012-10-10 2014-04-04 2014-08-05 2014-11-26	WO2006037038 AU2005289439 CA2581840 EP1797113 IL182225 IN2674/DELNP/2007 JP2008514203 US2009232841 US8101739 US2012156239	ALIVE ALIVE ALIVE ALIVE DEAD ALIVE ALIVE ALIVE ALIVE ALIVE			1 HNTPVYKLDISEATQVE 17 389 HNTPVYKLDISEATQVE 405	100.00

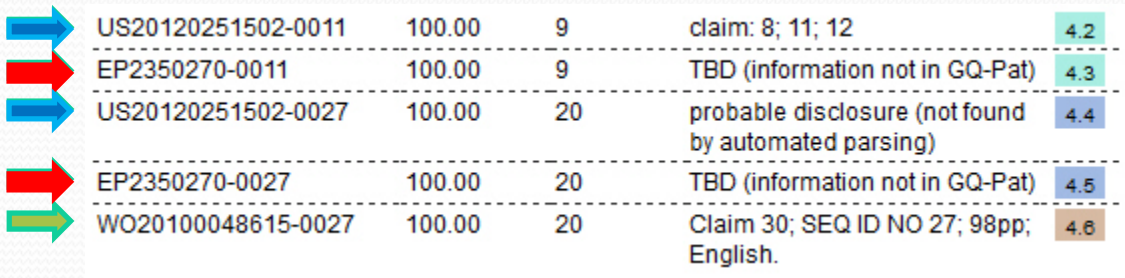
Working with Sequence Results

- Summarize sequence results for the family

	Title	Database	Patent Assignee	Query ID	Sequence Locations				
					Seq. ID Number	% Identity	Length	Location	
1.	PRODUCTION OF PEPTIDES IN PLANTS AS VIRAL COAT PROTEIN FUSION 1.1 Patbase	1.1 Patbase | link 1.2 GENESEQ | link	LARGE SCALE BIOLOGY CORP. 1.2 GENESE	query2	WO20050108564-0101	100.00	17	Example 6; SEQ ID NO 101; 115pp; English.	1.2
2.	Chimeric ebola virus envelopes and uses therefor 2.1 Patbase	2.1 Patbase | link 2.2 GPATPRT | link 2.3 GPATPRT | link 2.4 GPATPRT | link 2.5 GPATPRT | link 2.6 GENESEQ | link 2.7 GENESEQ | link	UNIV PENNSYLVANIA. 2.6 GENESE	query2 query3	US20050255123-0001	100.00	17	claim: 17	2.2
					WO03092582-0009	100.00	498	claim: 17	2.3
					WO03092582-0001	100.00	17	claim: 17	2.4
					US20050255123-0009	100.00	498	claim: 17	2.5
					WO20030092582-0001	100.00	17	Claim 17; SEQ ID NO 1; 107pp; English.	2.6
					WO20030092582-0009	100.00	498	Claim 17; SEQ ID NO 9; 107pp; English.	2.7
3.	ANTIGEN FRAGMENT AND TRUNCATION BASED ON EBOLA VIRUS ENVELOPE PROTEIN AS WELL AS APPLICATION 3.1 Patbase	3.1 Patbase | link 3.2 GENESEQ | link 3.3 GENESEQ | link	BIOENGINEERING RES INST ACAD MEDICAL SCI. 3.2 GENESE	query2	CN103864904-0008	100.00	17	Example 1; SEQ ID NO 8; 28pp; Chinese.	3.2
					CN103864904-0002	100.00	17	Example 1; SEQ ID NO 2; 28pp; Chinese.	3.3
4.	HUMAN EBOLA VIRUS SPECIES AND COMPOSITIONS AND METHODS THEREOF 4.1 Patbase	4.1 Patbase | link 4.2 GPATPRT | link 4.3 GPATPRT | link 4.4 GPATNUC | link 4.5 GPATNUC | link 4.6 GENESEQ | link	US DEPT HEALTH & HUMAN SERVICES. 4.6 GENESE	query7 query5	US20120251502-0011	100.00	9	claim: 8; 11; 12	4.2
					EP2350270-0011	100.00	9	TBD (information not in GQ-Pat)	4.3
					US20120251502-0027	100.00	20	probable disclosure (not found by automated parsing)	4.4
					EP2350270-0027	100.00	20	TBD (information not in GQ-Pat)	4.5
					WO20100048615-0027	100.00	20	Claim 30; SEQ ID NO 27; 98pp; English.	4.6

Working with Sequence Results

- Results may contain the same sequence from publications in the same family, or from different databases

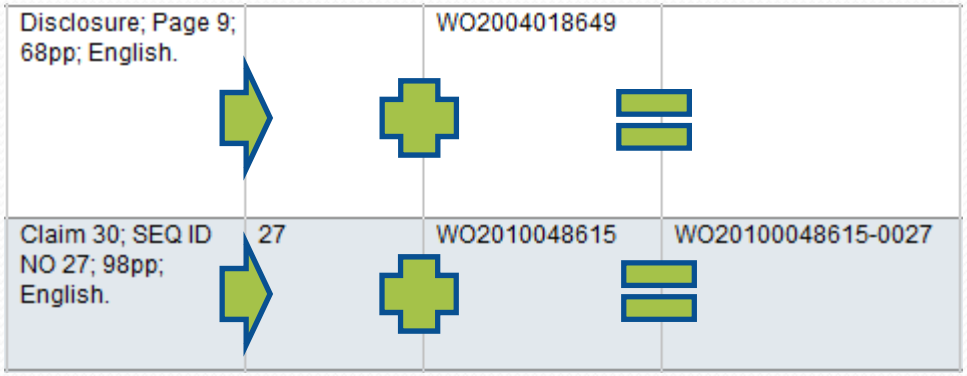


US20120251502-0011	100.00	9	claim: 8; 11; 12	4.2
EP2350270-0011	100.00	9	TBD (information not in GQ-Pat)	4.3
US20120251502-0027	100.00	20	probable disclosure (not found by automated parsing)	4.4
EP2350270-0027	100.00	20	TBD (information not in GQ-Pat)	4.5
WO20100048615-0027	100.00	20	Claim 30; SEQ ID NO 27; 98pp; English.	4.6

- Grouping by Sequence ID or Publication won't help in this case

Working with Sequence Results

- Automatically constructed “sequence ID number” when possible
  For example: GeneSeq



- Similar techniques used in GeneSeq, USGENE, PCTGEN

Working with Sequence Results

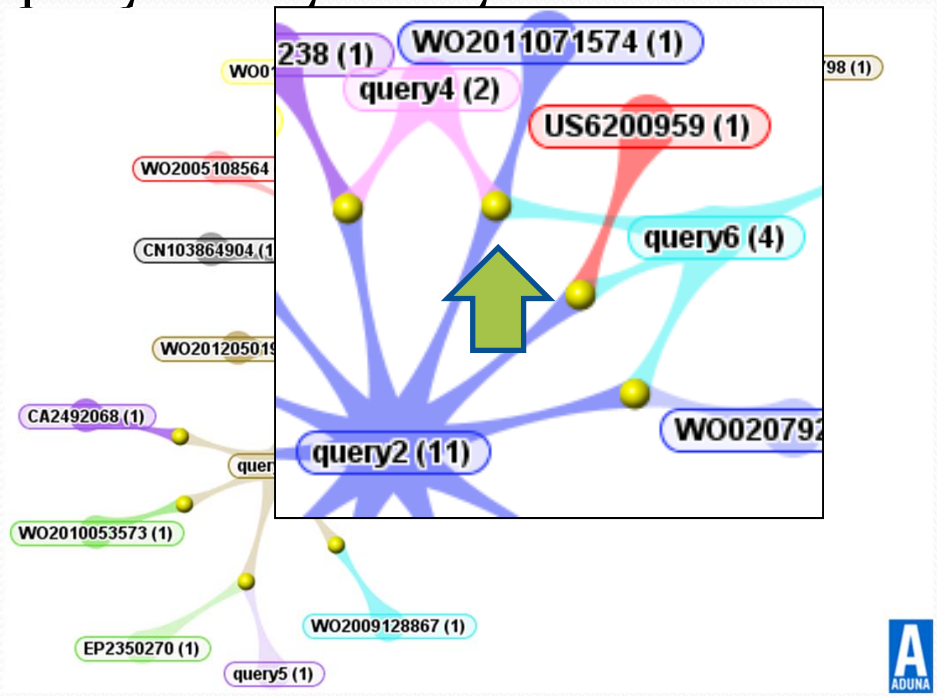
- Grouping by family makes sense when family databases are used
- Can we improve is the summary display?
- Idea: collapse sequences by SEQ ID NO within a family

US20120251502-0011	100.00	9	claim: 8; 11; 12	4.2
WO20100048615-0027	100.00	20	Claim 30; SEQ ID NO 27; 98pp; English.	4.6

- What about records without SEQ ID NO?
  Not collapsed, but still present

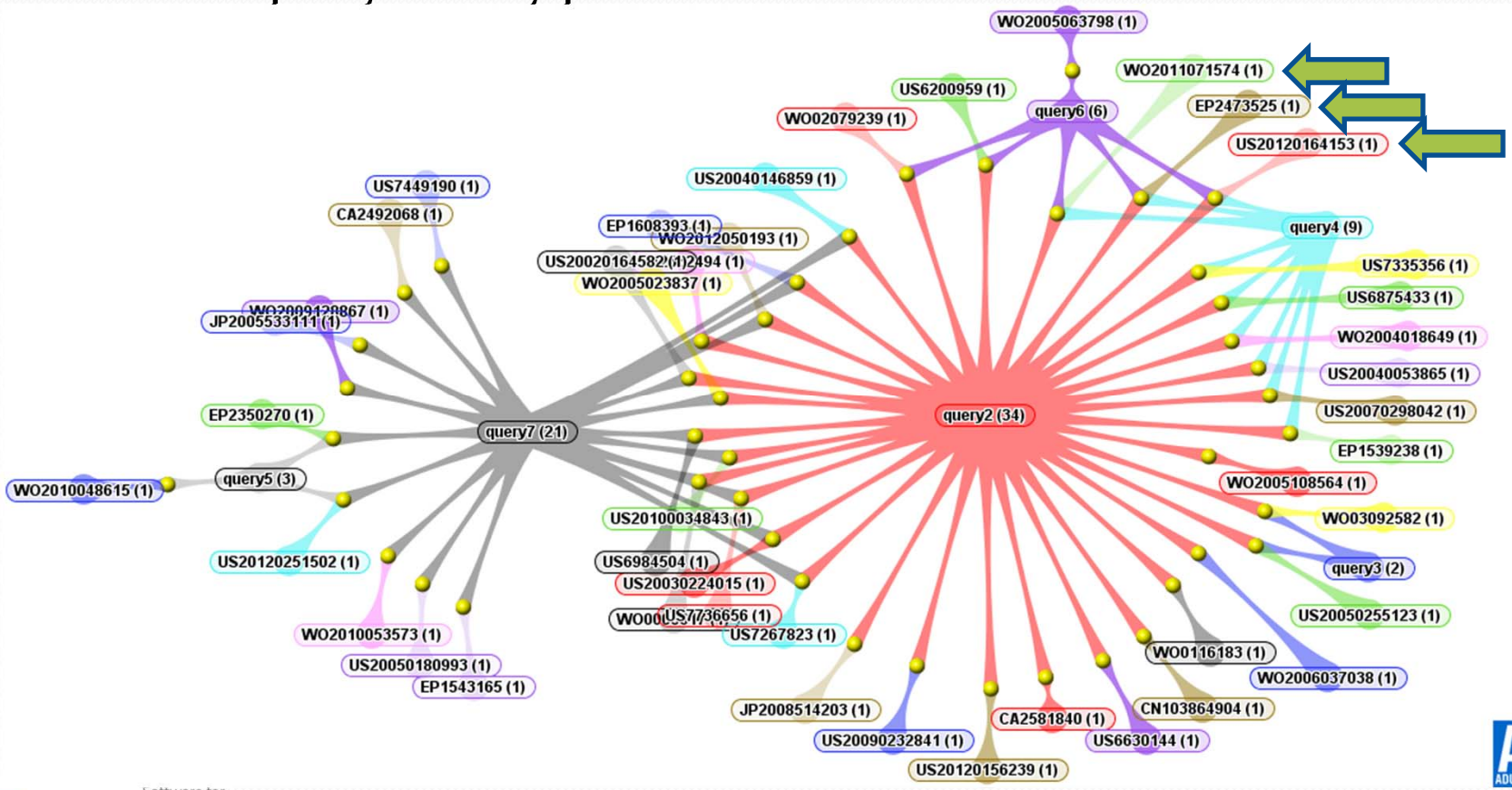
Working with Sequence Results

- Visualize query hits by family



Working with Sequence Results

- Visualize query hits by publication



Thank you!

www.bizint.com/slides