BIOINFO 2005

Proceedings of the 2005 International Joint Conference of InCoB, AASBi and KSBI

September 22-24, 2005 Busan, KOREA

http://www.ksbi.org/bioinfo2005



Doheon Lee
Limsoon Wong
Sangsoo Kim
Dae-Won Kim
Cheolmin Kim
Tin Wee Tan
Kwang-Hyung Lee

Hosted by

Association of Asian Societies for Bioinformatics Asia Pacific Bioinformatics Network Korean Society for Bioinformatics

Organized by

Electronics and Telecommunications Research Institute
Korea Basic Science Institute
Korea Institute of Science and Technology Information
Korea National Institute of Health
Korea Research Institute of Bioscience and Biotechnology
National Institute of Agricultural Biotechnology
National Livestock Research Institute
National Core Research Center for Systems Bio—Dynamics
Busan Genome Center
Chungbuk BIT Research—Oriented University Consortium

A Personalized Data Management System for Heterogeneous Biological Data Editing Kwang Su Jung, Sung Hee Park, Keun Ho Ryu	500
A Bridge Resistance Deviation-to-Time Interval Converter for Blood Pressure Measurement Chang-Soo Won, Seong-Hoon Kim, Won-Sup Chung, Sang-Hee Son	504
Semantic Content Browsing and Retrieval KyoungSoo Bok, JaeSoo Yoo	507
Roles of Conserved Serine Residues in Tobacco Acetolactate Synthase Eun Hye Park, Young Je Chung, Jung Do Choi	512
SPECIAL SESSIONS: National Research Institutes	
Excavation of ethnic single nucleotide polymorphism and its classification Sohyun Hwang, Seung-Woo Son, Sang Chul Kim, Young Joo Kim, Hawoong Jeong, Doheon Lee	521
AVSH: Annotation and Visualization System for Haplotype map Jungsun Park, Tae-Hui Hong, Hyo Jin Kang, Jinhee Jung, Taewook Kang, Sangcheol Kim	521
REPEATOME for Primate: an integrative analysis database of primate repeat elements	
Taeha Woo, Jung Min Seo, Tae-Hui Hong, Byung-Chul Kim, Ungsik Yu, SangSoo Kim, Chang-Bae Kim	522
KUGI: a database and search system of Korean UniGene and Pathway Information Jin Ok Yang, Yoonsoo Hahn, Nam-Soon Kim, Ungsik Yu, In-Sun Chu, Yong Sung Kim, Hyang-Sook Yoo, Sangsoo Kim	, 522
Gene Expression Analysis: GEDA-C and GEDA-Diag	523
Ho-Youl Jung, Ji Eun Kim, Seon Hee Park Protein Structural Data Mining	
Chan Yong Park, Sung Hee Park, Dae Hee Kim, Seon Hee Park	523
bioINET: bio-object Inter-relationship NETwork system	
Jae Hoon Choi, Jong Min Park, Seon Hee Park	524
Integrated Bioinformatics System	
Sang Joo Lee · · · · · · · · · · · · · · · · · ·	524
The prosperity of GBIF and its effort to KBIF's Activities	525
Sungsoo Ahn, Hyung-Seon Park	323
Function Annotation Research Model	525
Jongsun Jung	
Chan Park	526
In Silico Functional Analysis of Disease Genes	
Kuchan Kimm · · · · · · · · · · · · · · · · · ·	. 526
Overview of Agricultural Bioinformatics Projects by IMT-2000 Special Grant - Development of Bioinformatics Application Systems for Agricultural Bioresources	
Jang-Ho Hahn, Yong-Hwan Kim, Chang-Kug Kim, Gang-Seob Lee, Eun-Gi Cho, Kunsoo Park,	527
Hae-Jine Kim, Hyun-Suk Park, Jae-Won Lee, Gil-Bok Lee	-
The Development of Core Collection Program using Heuristic Algorithm for Plant Genetic Resou	. 527
Hoon-Gi Jung, Jung-Hoon Kang, Jae-Gyun Gwag, Kyung-Ho Ma, Kyu-Won Kim, Yong-Gyu Park	0 200
Mapping OF >580,000 Expressed sequence tags from full-length cDNA clones from japonica rice updating the rice oligomicroarray system to cover the whole expressed genes in rice	
THE RICE FULL-LENGTH CONSORTIUM	• 528
A Data Warehouse for Visualizing Bio-pathways of Agricultural Organisms	. 528
Tryum beek ram	
Developing A Fragment Assembly Program Jin Wook Kim, Inbok Lee, Joong Chae Na, Kangho Roh, Sunho Lee, Kunsoo Park	

GO

Ge

Ex

Str

Sn

A Personalized Data Management System for Heterogeneous Biological Data Editing

Kwang Su Jung

Sung Hee Park

Keun Ho Ryu

Database/Bioinformatics Laboratory, Chungbuk National University, South Korea Email: {ksjung, shpark, khryu}@dblab.chungbuk.ac.kr

ABSTRACT: The biological sequences in the biological laboratories have been produced since the techniques to get the sequences of genomes or proteins in HGP(Human Genome Project) have been improved. Unfortunately, there are scarcely the software packages to deal with the sequential data in most of biological laboratories and they are just stored in file formats. The integrated management system of biological data is required to manage the sequence data taken from other open databases to improve the analysis of the sequence data in our biological labs. We therefore suggest the system to edit, store, search biological information, and convert the formats of the sequence data, as well as to integrate and manage the data.

1 INTRODUCION

The biological Sequences have been made in many laboratories since the techniques to get the sequences of genomes or proteins in HGP(Human Genome Project) have been improved. The biologists connect the public biological Databases[1-4] and retrieve sequences which is similar with what they have, then this work is utilized in homology research, functional analysis and prediction. Unfortunately, there are scarcely the software packages to deal with the sequence data in most of biological laboratories and they are just stored in file formats

The integration and management technique of heterogeneous sequence data from public sequence databases is widely used to make diverse information and prediction. Thus the database management technique that is suitable for a sequence data is required. Especially, an integrated data model which handles the modification of program and data is needed for analysis on the various programs.

In this work, we support editing and converting among heterogeneous public biological database flat files as well as own sequences in laboratories. The BSML(Bioinformatic Sequence Mark up Language)[5] based on XML is utilized for representing the complex and hierarchical biological data. The biologists easily analyze their accumulated biological data using our system and apply the results of biological data analysis to medical science and pharmacy.

2 RELATED WORKS

2.1 BSML

The BSML is one of open standard of XML data in

bioinformatics research. BSML represents good bio-physical features as well as visualization of biological sequence data comparing with other languages such as abstract annotations. It also supports converting formats among biological flat files base on flat-form independence. In this point, BSML is more complete and specifiable than other XML formats in bioinformatics. BSML which our system supports is used in the area of representing gene research results and biological molecules, and exchanging them, hence many laboratories and company choose BSML as new standard.

2.2 Genomic Workspace

The Recent tools for converting formats among standards are implemented by JAVA or Perl module. One of these worthy converting softwares is Genomic Workspace[6]. It converts Genbank[1] flat file into BSML format and includes Genomic Viewer for a good graphic visualization of biological information.

2.3 Staden Package

The Staden Package[7] is developed for managing and analyzing the sequence from sequencing machine by Medical Research Council Laboratory in U.K. It constitutes several tools as following. *Trev* represents raw experimental file from sequencing machine. *Trace_diff* describes the mutation information of reference and Trace data. *GAP4* edits contigs and assembly of sequences. Pregap4 is used for pre-processing the assembly data. Spin analyzes the result sequence such as retrieving similar sequences and some other operations.

The Staden Package analyzes and simply manages sequences in sequence experimental files but doesn't store sequences into database and retrieve. It can't process a mass of sequences.

2.4 Sequence Data Formats

Each flat file from public biological database has different format. Genbank[1], Swiss-Prot [2], PIR-Codata[3] and PDB[4] flat files are generally used. Also, each sequence analyzing software uses different format but FASTA[8] established by Pearson is the most common format.

> SQ2002060300001 GGTACCTTCTGAGGCGGAAAGAACCAG CCGGATCCCTCGAGGGATCCAGACATG CTTACCGGATACCTG

Figure 1: An example of FASTA format

- 500 -

are released are can different Text

excl XM re-u repr Bio

biol

XM

3

Figi brie

Biok

3.1

The each from Part extends flat

BS bet ele

ext

3.1 Bi

da Ed The problems of biological formats I mentioned above are following. The formats can be modified when they are released. To understand the range of field and field value are difficult and data types in the same field in each format can be different. The conversion for different formats needs different parsers to extract the interesting field.

Thus applying XML is essential to visualize and exchange the sequence data and the bio-molecules because XML has high level format conversion technology and re-usability. There are a lot of markup languages for representing the genome data such as BSML, BioML(Biopolymer Markup Language)[9]. The public biological databases[1-4] are also trying to produce the XML based formats for representing biological information.

3 SYSTEM ARCHITECTURE

Figure 2 shows the system architecture we suggest. We briefly explain each component of our system.

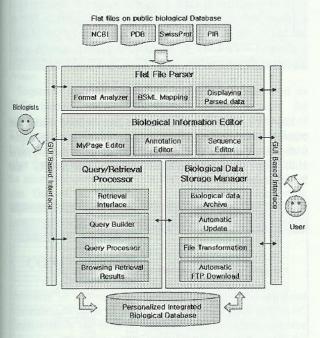


Figure 2: System Architecture

3.1 Flat File Parser

The Flat File Parser extracts meaningful data fields from each flat file in order to integrate heterogeneous formats from the various public biological databases. The Flat File Parser which is suitable for each flat file is developed for extracting the interesting fields from divergent biological flat files, then the integrated BSML model is created using extracted data. We developed Flat File Parsers per each flat file such as Genbank, Swiss-Prot, PIR-Codata, FASTA, BSML and so on. Figure 3 shows the mapping information between the parsed interesting fields and BSML elements(attributes) for loading on the system.

3.2 Biological Information Editor

Biological Information Editor edits annotation and sequence data. Our Editor consists of MyPage Editor, Annotation Editor and Sequence Editor.

MyPage Editor supports to make user defined formats that constitute the interesting fields extracted from various flat files such as Genbank, Swiss-Prot, PDB, PIR-Codata, BSML, FASTA, data on the local database and so on. Once MyPage format is made by user, Modification, insertion and deletion of the field in the format are also performed by MyPage Editor. The user defined format through MyPage Editor is stored as an XML format on the disk or local database.

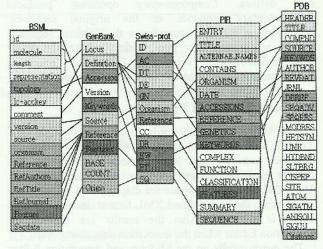


Figure 3: Mapping Information among biological Formats

While MyPage Editor creates user-defined formats and performs insertion, modification and deletion of the interesting field extracted from each format, Annotation Editor performs the same operations for annotation data in each format. Therefore original source(flat file) is maintained their own frame. The modified original source stored as BSML format on the disk and local database if the user wants. Annotation Editor mainly handles annotation data of interesting sequence such as types, taxonomic information, citations, reference data, authors, specific sites of sequences and so on.

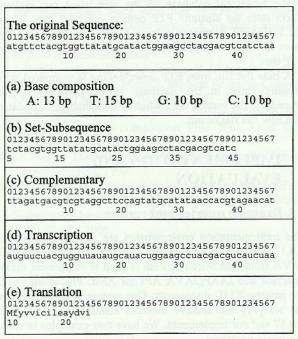


Figure 4: Sequence operations

Sequence Editor[10] supports the useful operations handling sequence, so it produces a new sequence from original sequence. We will briefly introduce several operations on Figure 4. Base Composition operation calculates the composition of DNA(Deoxyribo-Nucleic Acid) and shows us the percentage and the numbers of each Nucleotide(Adenine, Guanine, Cytosine and Thymine). Set-Subsequence operation makes a new sequence from original sequence using starting and ending index which the defines. Complementary operation produces user complementary sequence of the original sequence. Transcription operation transforms the original sequence which is proven as gene, into mRNA(Messenger Ribo-Nucleic Acid) sequence. Translation operation transforms mRNA sequence into Protein sequence.

3.3 Query/Retrieval Processor

The biologists input keywords to retrieve through GUI (Graphic User Interface) environment and Query/Retrieval Processor retrieves the local database and returns the suitable results for retrieval options users selected such as AND, OR. The query results are displayed on the MyPage Editor when the user defined XML format through MyPage Editor is retrieved. Also the results are showed on Annotation Editor when the annotation data is retrieved.

The query results on the MyPage and Annotation Editor is also modified by adequate Editors if user wants.

3.4 Biological Data Storage Manager

Sequence data and annotation data for sequence is separately stored because the fast accession to biological data and retrieval are needed in order to analyze data. When biologist wants to make other format from the database, for example FASTA format, the necessary biological data on the database for making a new format is retrieved.

Our database maintains up-to-date biological information through automatic FTP (File Transfer Protocol) download and update. Users can monitor new release information of public biological database on the web using our system. Users store the absolute FTP path in our FTP lists. When the biological databases release a new version, what the users have to do is just choosing appropriate FTP path in FTP lists.

A new flat file from FTP downloading is compared with existing entry in the database to be updated. In the procedure, the system asks whether the database is updated as a new flat file or not.

4 IMPLEMENTATION AND EVALUATION

4.1 Example queries and results

The implementation environments are following. MySQL 4.0.17 is used to store genome sequence and annotation data. JAVA from Sun Mirco System is used to implement major interface and JAXP(JAVA API for XML Processing) v1.2 for parsing XML documents. Our system is distributed on the web[11] and archive contains install files of MySQL 4.0.17 and our implements. We have two examples to show our implementation.

EXAM 1. User wants to make Complementary, Transcription, Translation, Set-Subsequence operated Sequence of original sequence in Genbank locus name 'DRONCX' with base composition. The upper window on Figure 5 shows the original sequence and the lower one shows the derived sequences from operations.

AUR	CAG	UUG CUA	CUC		AAA	UCS	AUA	UCC	ACC
ACC	GCC	CAG	UCG		AUC	AAA	GUE	CAS	GCC GAG
ACA	GAA	GC6	CGA		8CG	340	CUE	AAU	GUU
ACC	UCC	UCC	AGC		ABU	AAE	ac	AGC	CAG
GAC	GAU	OGC	CAC		CUC	ABC	099	CGA	CUA
AGG	CAG	SUC	AGC	CAU	880	BAB	GAS	GGC	6AC
GAG	GGC	GCG	ECG		CAA	ALIS	GAE	GAC	GAG
ig.	GAA	CAG	AUG	ACC	AAG	GUG	CALL	GGA	GAA
BCA	008	GAC	aca	BAG	848	CUB	CSC	GAA	UGU
F	일본서 DNA역 Rotatio SetRan	u 정보서	ā	cceatos ATTOCT elgoagtij elgoagtij	ctcctte	eatogat	FFAAA afteacc	CTTTTA tgcgcex	
MUD (00) (MUL)	ALAKA CARACA	COLUMN TO THE REAL PROPERTY.	25	AUGCAC	LIUGCI	JCCUL	LAALIO	CALLALI	BIC SEASON STATES
	Same	CUE	SALES CONTRACTOR	MOLLLIK		~~~~~	***************************************	*********	Manager & Commission of the Co

Figure 5: Sequence Editor

EXAM 2. User wants to create a user defined format that consists of interesting field from various file formats. Figure 6 shows creating a user defined format from Swiss-Prot(id 100K_RAT), Genbank(locus name DRONCX) and PIR-Codata (entry CCHU).

pp age	州智慧里思	집 주석정보편집	경색 네이트	테베이스 도움임				10000	
4	.		*******						
240	859	052	A Arms	****	80.050				
100	Locus	SHONCX	at these	2,000,82900	ORDONARO		eased, ber-Coreda Ce	DICAGE!	
1-1-		100M PAT				カウ	養定額	952	Concents
3-30-	ENIRY	COU Mypec					rock	DRICHCK 3176	Control of the last of the las
SHARK	FT	DOMAN 7					DEFINITION	Crossorale volenogester	
5-300-	KEYWORDS	acelythese eve		·		15	ACCESSION	L39036	
100	CURNAL	Am. J. Physici.		· nonnomes	·····	100	MERSION	L39635.1 G11905801	Wile Williams
S. bear	EXPRIAL	Am J. Physiol.				3	PEYWORDS	: Merc's exchanger, antipor	
30 3	(Plant)	Construction .		***************			SOUNCE	Dropophie resergester	
3000	SEQUENCE	MGOVENSHIEL		***********		16	ORGANISM	Dropophile malanogester	described to
2.30	960	MARSANCOAN.		·			rerenexe	(1 (Leases 1 to 3178)	
3000		*******	William Control	dennerous co	ere area area area	II.	AUTHORS	SIGNATE DH , KONAP V	
***	*****					11	me	Alternative spacing of the	THE REAL PROPERTY.
5333			*******		*******	£C.	TOTANT.	Arm. N. Y. Acad. Sci. 770	
1000	*****						NEUTHE	96250000	
1883			****			1	PLEMED	000000	- 77777 1 1 1 2 2 2 2
1888	******		******		******		PETERENCE AUTHORS	2 (bears 1 to 2176)	
1000			*****				AUTHORS .	PLANLEN, A. VARIVAS.	
1888	*****				*****	i.C.	≣TLE .	Mo+Ca2+ extrenge in D	
1333			******				JOURNAL .	Art. J. Physiol. 273 (1 Pt 1	
1000	******		233434				MEDUNE	67096273	
\$333		********	* * * * * * * * * * * * * * * * * * * *			Between	PLEMED	1252464	
2000						LE.	REFERENCE	(3 fberes ! to 3176)	
1000	*******	********	3000000			15.	AUTHORS	SOMOODH	
1000	3000000		******	*******	********	15	IIILE	Desci Submeson	COLUMN TO SERVICE
2323		*********					JOL PHAL	(Submitted (19-MAR-1997)	
222				******		88C.	COMMENT	Con Mar 24, 1007 this sec.	
888		*********				1	FEATURES	!Ltcstor/Quettery	
200			******			2.4-0	8040e	11.3175 torgonism-Tropo	***************************************
5000	*****	*********			*******	1	gere	148-2601 /pro-79CF	
2000						1100	cos	45 2901 /gene-1907 Au	L-000000000000000000000000000000000000
200,000,000	**********		0.4.6.0.6.0.00		0.0000000000000000000000000000000000000	508MW	LUCIO DE LO CONTRACTOR DE LA CONTRACTOR DE	Market and a second second second	THE PERSON NAMED IN COLUMN

Figure 6: MyPage Editor

4.2 Evaluation

In this paper, we implement new issues that we mentioned in introduction with deep consideration. The evaluation of the proposed system is performed according to comparing with Staden Package, Genbank, Genomic Workspace and so on. Table 1 shows the result of comparing.

Storing the historical data of sequence release version is applied for analyzing the change of protein sequence, we expect that new useful protein and drug can be discovered based on analysis of version sequence.

So Mai Fil So So Vo

Creating and analysi purpose of a huge numbe enough star complement and a know sequence oper Primer.

Sequence
XML forms
supporting es
XML represe
Definition), s
other formats
the most com
uses. One of
sequence from
format.

5 CONC

According to information, a Characteristics sequences are need of manag should reflect I most current bi as repositories:

Therefore, the of genome so sequences for no includes format for collected no and handles see BSML based of extract data fit formats. To ma management system automated database using F term study of bid to accumulate the biological labora

mplementary, ce operated locus name r window on le lower one



format that nats. Figure wiss-Prot(id ICX) and



mentioned aluation of comparing ace and so

version is uence, we discovered

Item	Staden Package	Genbank	Genomic Workspace	Proposed System	
Sequence Manipulation	Support	None	None	Support	
File Format EMBL forma		Genbank/ASN.1 /BSML	BSML	XML Format based on BSML DTD	
Storing File Sequence		DBMS	BSML Documents/ Text	DBMS	
Sequence Versions	None	Support	None	Support	
File None Transformation		None	Support	Support	

Table 1: Comparison with existing Sequence Management Systems

Creating and designing a new sequence for experiment and analysis through Sequence Editor is possible. The purpose of a PCR (Polymerase Chain Reaction) is to make a huge number of copies of a gene. This is necessary to have enough starting template for sequencing. In this work, a complementary sequence of one of double helix is needed and a known *Primer* is required to synthesize. Therefore sequence operations we suggest are usable enough to design *Primer*.

Sequence information in this system is represented as XML format and converted into other formats for supporting efficient sequence analysis. That is, because of XML representation based on BSML DTD(Document Type Definition), sequence information is easily converted into other formats used in various analysis program. FASTA is the most common format that the most analyzing software uses. One of the advantages of our system is converting sequence from database or various sources into FASTA format.

5 CONCLUSION

According to improving techniques to get the biological information, a mass of biological data have been produced. Characteristics of these biological data including genome sequences are heterogeneous and various. Although the need of management systems for genome sequencing which should reflect biological characteristics has been raised, the most current biological databases provide restricted function as repositories for biological data.

Therefore, this paper describes the management system of genome sequences and annotation data of those sequences for manipulating the formats in bioinformatics. It includes format transforming, editing, storing and retrieving for collected nucleotide sequences from public databases, and handles sequence produced by experiments. It uses BSML based on XML as a common format in order to extract data fields and transfer heterogeneous sequence formats. To manage sequences and their changes, version management system for originated DNA is required. This system automatically updates the information in local database using FTP. This research is widely used for a long term study of biology, medical science, pharmacy and helps to accumulate the own sequence information at the level of biological laboratories.

ACKNOWLEDGMENT

This work was supported by the Regional Research Centers Program of Ministry of Education & Human Resources Development in Korea.

REFERENCES

- [1] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler. GenBank. Nucl. Acids. Res, 30:17--20, 2002.
- [2] A. Bairoch and R. Apweiler. The Swiss-Prot protein sequence database and its supplement TrEMBL in 2000. Nucl. Acides. Res, 28:45-48, 2000.
- [3] C. H. Wu, L. L. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, R. S. Ledley, P. K. Baris, E. Suzek, C. R. Vinayaka, J. Zhang, C. Winona. The Protein Information Resource. Nucl. Acids. Res, 31:345--347, 2003.
- [4] H. Berman, M. Westbrook, J. Feng, Z. Gilliland, G. Bhat, T. N. Weissig, H. Shindyalov, P. E. Bourne. The protein data bank. Nucl. Acid. Res, 28:235--242, 2000.
- [5] J. Spitzner, Bioinformatics Sequence Markup Language Manual, LabBook Inc., 1997.
- [6] http://www.rescentris.com/
- [7] J. Bonfiled, K. F. Beal, M. Jordan, Y. Cheng, R. Staden. The Staden Package Manual. Medical Research Council Labortory of Molecular Biology, 2001.
- [8] W.R. Pearson, D.J. Lipman. Improved tools for biological sequence comparison. Proc. Narl. Acad. Sci, 85:2444-2448, 1988.
- [9] D. Fenyo, The Biopolymer Markup Language, Oxford University Press, 1999.
- [10] S. H. Park, K. S. Jung, K.H. Ryu. Implementation of an Information Management System for Nucleotide Sequences based on BSML. KISS Journal D, 32(1):24--42, 2005.
- [11] http://dblab.chungbuk.ac.kr/~kistep2004