

# Introduction to web based annotation system for Structural-Biological Whole Cell Project *Thermus thermophilus* HB8

## *Thermus thermophilus* HB8用に開発された Web baseのアノテーションシステムの利用方法

Akinobu FUKUZAKI <sup>1,2</sup>、Fumikazu KONISHI <sup>1,2</sup>、Takeshi NAGASHIMA <sup>1</sup>、Kaori IDE <sup>1,2</sup>、  
Mariko HATAKEYAMA <sup>1,2</sup>、Seiki KURAMITSU <sup>2,3</sup>、Akihiko KONAGAYA <sup>1,2</sup>

福崎 昭伸 <sup>1,2</sup>、小西 史一 <sup>1,2</sup>、長嶋 剛志 <sup>1</sup>、井手 香 <sup>1,2</sup>、  
畠山 眞里子 <sup>1,2</sup>、倉光 成紀 <sup>2,3</sup>、小長谷 明彦 <sup>1,2</sup>

( <sup>1</sup>Bioinformatics G., RIKEN GSC <sup>2</sup>RIKEN Harima Institute <sup>3</sup>Grad. Sch. of Sci., Osaka Univ.)

( <sup>1</sup>理研 GSC ゲノム情報科学研究G <sup>2</sup>理研 播磨 <sup>3</sup>大阪大学 理学研究科 生物化学)

e-mail: akki@gsc.riken.jp

高度好熱菌丸ごと一匹プロジェクトでは構造機能解析をはじめとして系統的機能解析のためのトランスクリプトーム解析やプロテオーム解析、メタボローム解析、システム生物学的研究を多数の研究者の協力のもと進めている。その中でプロジェクト参加者の間で実験に必要なゲノム配列や遺伝子情報、タンパク質の情報を共有する場を提供するとともにそれらの情報を更新できるようなシステムの必要性がもたられてきた。我々はWeb技術とデータベース技術を用いてプロジェクトに参加している研究者達が便利に活用できるようなアノテーションシステムを指向して開発を続けてきた。今回は連携研究会の場をおかりして我々が開発したシステムの使用方法の紹介を行う。

### アノテーションシステムでできること

本システムでは現在のところ以下のような機能が利用できる。これらの機能は今後も拡張や追加がなされるのでより新しい説明はアノテーションシステムのWebサイトにあるオンラインドキュメントを参照していただきたい。

#### 配列情報の検索

TTナンバーでORFを検索する

キーワードで検索する

PDBで検索する

配列のホモロジーで検索する

#### アノテーションの更新

TTナンバーの対応付けの更新

機能名の更新

PDBの対応付けの更新

## システムへのアクセス

アノテーションシステムはユーザー認証システムによって保護されたWebサイトに設置されている。よってシステムへ接続するにはあらかじめ申請手続きをしてユーザーアカウントを取得する必要がある。本システムへのアクセスにはWeb Browserを利用するが、IEやMozilla等の一般的なWeb Browserで利用が可能である。なるべく最新版を利用することをおすすめする。

<https://access.obigrid.org/thermus>へアクセスすると以下のようなユーザー認証の画面が表示される。ここで取得したユーザー名とパスワードを入力しLOGINボタンをクリックする。



Access.obigrid.org ログイン

ユーザー名

パスワード

新規ユーザ登録は[こちら](#)からどうぞ。

 Access.obigrid.org is fully contributed by **RIKEN-GSC**  
Copyright(C) 2002-2004 OBIGrid, All Rights Reserved.  
If you have any question, please contact: [info@obigrid.org](mailto:info@obigrid.org)

正常にログインできると以下のようにポータルサイトのFrontPageが表示される。ポータルサイトではアノテーションシステムへのリンクだけでなくオンラインドキュメントやメンテナンス情報、さらにはユーザーが自由に使えるMy pageなどが提供される。ここではアノテーションシステムの説明を行う。アノテーションのメニューページへいくにはContentsのところのAnnotation Systemをクリックする。



**FrontPage**  
<https://access.obigrid.org/thermus/2FrontPage>

[ Reload ] [ New | Edit | Freeze | Diff | Upload ] [ My page | Front page | List of pages | Search | Recent changes | Backup | Help ]

recent(10)  
2004-06-24  
• akki/comments  
• akki  
• akki\_demo/com  
• akki\_demo  
• FrontPage  
• Annotation System  
• test  
• MenuBar  
• TextFormatting  
• links

**TTHB8 portal**

**maintenance infomation**

This server will have a maintenance from 15:00 to 17:00 every day.  
Annotate operation during the maintenace can be damage annotation data!!  
Be sure it is not a maintenace time before you work on the system.

**Contents**

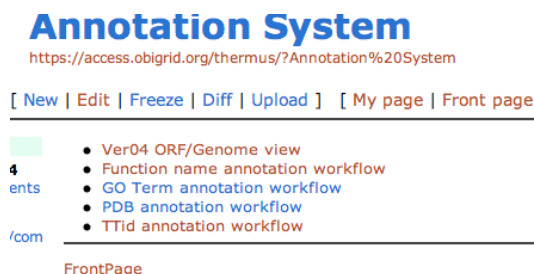
- Annotation System

Last-modified: Thu, 24 Jun 2004 18:03:07 JST (23h)  
Link: akki(20h) akki\_demo(20h) Annotation System(23h) test(23h) links(1d)  
Contact person:  
Mariko Hatakeyama and Fumikazu Konishi  
Bioinformatics group, Genomic Sciences Center, RIKEN



## アノテーションシステムのメニュー

アノテーションシステムはいくつかのメニューを持っている。現在のバージョンでは情報を閲覧するviewが1つとアノテーションを行うworkflowが4つある。これらのメニューをクリックすることでアノテーションシステムの各機能へアクセスすることができる。以下に各機能の説明をする。



### Ver04 ORF/Genome view

情報の閲覧を目的としている場合はこの機能を使用する。ORFをベースとした各種検索とアノテーション情報や配列情報、ゲノムのLocusの情報が提供される。

### Function name annotation workflow

ORFの機能名を更新する時に使用する。

### GO Term annotation workflow

GO Termの対応付けを更新する時に使用する。

### PDB annotation workflow

PDBの対応付けを更新する時に使用する。

### TTid annotation workflow

TTナンバーの対応付けを更新する時に使用する。ほとんどのTTナンバー-ORFは最新のゲノム配列に対して位置を特定できており対応するORFも判明しているが、一部の例外がある。もし調べたいTTナンバーを検索しても見つからない場合はこちらで情報を更新する必要がある。

## 実際の使用例

ここからは実際にシステムを使った例を示す。

TTナンバーでORFを検索する

キーワードで検索する

PDBで検索する

Ver04 ORF/Genome viewを使って各種条件で検索を行い、該当するORFを呼び出すことができる。まずAnnotation SystemのメニューでVer04 ORF/Genome viewをクリックする。すると右のような閲覧画面が表示される。

赤枠で囲まれたところが各種表示画面の切り替えタブで、クリックすることで切り替えることができる。検索をしたい場合はRegionSearchをクリックする。するとタブの下ツール表示部分が検索用の画面に切り替わる。

青枠で囲まれたところがRegionSearchツールの中でさらに機能を選ぶタブである。ここではまずSimpleを選ぶ。SimpleではTTナンバーやPDBのID、機能名のキーワードなどを入力して検索することができる。

例えばTT0051とKeywordに入力してSearchをクリックすると2件の結果が表示される。TTHB84orf1388300というRegionはゲノムのForward strand側にあるORFで、もう一方のTTHB84tt\_5005100はTTナンバーORFをこのシステムのデータベースに登録する時に付けられたIDである。このようにTTナンバーORFとゲノムのORFの両方が見つかった場合はそのTTナンバーの対応付けがきちんとされていることになる。検索結果のTTHB84orf1388300をクリックするとそのORFの情報を表示するモードに切り替わり、例えばHomologのnr/ntタブをクリックすることでBLASTによりどのような配列がヒットしたのかを確認することができる。ここでgi:のところをクリックするとNCBIのサイトへジャンプしてそのエントリーの詳細を見ることができるし、ヒット位置を示すバーをクリックするとBLAST2Sequenceによる詳細なアライメントを見ることができる。

さらにPDBのIDで検索する時は1IW7のようにPDB IDをkeywordに入力することで該当するORFを見つけることができる。この場合は3件ヒットした。

また、DNA repairのようにスペースを間にいれて複数のキーワードを入力することでそれらの全てのキーワードが機能名に含まれるORFを見つけることができる。この例ではDNA repair protein関係のORFだけでなくDNA mismatch repair proteinも見つかっている。

TTTHB8 Region Information Viewer TTHB84orf1388300

| reg_id           | name                           | synonym           |
|------------------|--------------------------------|-------------------|
| TTHB84orf1388300 | putative threonine dehydratase | TT0051 GO:0005520 |

RegionSearch Annotation Annotation History Homolog Data RegionView Genome Browse

RegionSearch

Simple BLAST Advance RevHomolog File

Keyword: TT0051  
☒ ORFs on Genome ☒ TT ORFs  
search

View results by TSV

1 entries:

| reg_id           | ttid   | ann_name                       |
|------------------|--------|--------------------------------|
| TTHB84orf1388300 | TT0051 | putative threonine dehydratase |

TTTHB8 Region Information Viewer TTHB84orf1388300

| reg_id           | name                           | synonym           |
|------------------|--------------------------------|-------------------|
| TTHB84orf1388300 | putative threonine dehydratase | TT0051 GO:0005520 |

RegionSearch Annotation Annotation History Homolog Data RegionView Genome Browse

RegionSearch

Simple BLAST Advance RevHomolog File

Keyword: TT0051  
☒ ORFs on Genome ☒ TT ORFs  
search

View results by TSV

2 entries:

| reg_id           | ttid   | ann_name                                       |
|------------------|--------|------------------------------------------------|
| TTHB84tt_5005100 | TT0051 | threonine dehydratase (deaminase) (catabolic?) |
| TTHB84orf1388300 | TT0051 | putative threonine dehydratase                 |

TTTHB8 Region Information Viewer TTHB84orf1388300

| reg_id           | name                           | synonym    |
|------------------|--------------------------------|------------|
| TTHB84orf1388300 | putative threonine dehydratase | GO:0005520 |

RegionSearch Annotation Annotation History Homolog Data RegionView Genome Browse

TTM ORF nr/nt PDB RefSeq

Current Annotation  
Name: putative threonine dehydratase GO Term: GO:0005520

☐ pos\_start ☐ pos\_end ☐ percent\_identity ☐ hit\_len ☐ e-value ☐ score (Option) 5 3

TTHB84orf1388300 936 (bp)

BLASTP

| gi          | hit                                                                                                              | description |
|-------------|------------------------------------------------------------------------------------------------------------------|-------------|
| gi:21219342 | putative threonine dehydratase [Streptomyces coelicolor A3(2)]                                                   |             |
| gi:15597979 | probable acetylthreonine dehydratase, degradative [Pseudomonas aeruginosa PA01]                                  |             |
| gi:8322831  | Pyridoxal-phosphate-dependent enzyme homologous to mouse glutamine racemase; YADP Rep [Saccharomyces cerevisiae] |             |
| gi:23128281 | COG1171: Threonine dehydratase (Pilotoc punctiforme)                                                             |             |
| gi:17549735 | PROBABLE AMINO-ACID DEHYDRATASE PROTEIN [Pastorina solaracearum]                                                 |             |

BLASTN

Keyword: 1IW7  
☒ ORFs on Genome ☐ TT ORFs  
search

View results by TSV

3 entries:

| reg_id            | ttid   | ann_name                                            | PDB  |
|-------------------|--------|-----------------------------------------------------|------|
| TTHB84orf11036500 | TT0791 | RNA polymerase principal sigma factor               | 1IW7 |
| TTHB84orf11922000 | TT0781 | DNA-directed RNA polymerase beta' chain (RpoC) 1IW7 |      |
| TTHB84orf11922000 | TT0780 | DNA-directed RNA polymerase beta chain (RpoB) 1IW7  |      |

Keyword: DNA repair  
☒ ORFs on Genome ☐ TT ORFs  
search

View results by TSV

7 entries:

| reg_id           | ttid   | ann_name                         | PDB |
|------------------|--------|----------------------------------|-----|
| TTHB84orf1038200 | TT0726 | DNA repair protein RadA          |     |
| TTHB84orf1057400 | TT1655 | DNA repair protein RecO          |     |
| TTHB84orf1194100 | TT0735 | DNA mismatch repair protein MutL |     |
| TTHB84orf1194100 | TT0734 | DNA mismatch repair protein MutS |     |
| TTHB84orf1218000 | TT0736 | DNA repair protein UvrA          |     |
| TTHB84orf1234900 | TT0755 | DNA repair protein RecN          |     |
| TTHB84orf1241200 | TT0738 | DNA repair protein UvrC          |     |

配列のホモロジーで検索する

配列のホモロジーでORFを探索することもできる。例えば e.coliのLactose Operon repressor (P03023)をNCBIのサイトからとってきて検索してみる。RegionSearchツールのBLASTタブをクリックすると配列を入力する画面になる。ここにP03023のペプチド配列をペーストし、Sequence TypeをPeptideにセットしてから、searchをクリックする。

するとこの配列と似た配列を持つORFがいくつか表示される。このように配列のホモロジーでORFを探す機能もあるので機能未知の配列であったり、まだアノテーションがきちんとされていない配列であってもゲノム上のどのORFに近いものかを見つけることができる。

[illegible]

## TTナンバーの対応付けの更新

## 機能名の更新

## PDBの対応付けの更新

アノテーションデータはまだ多くの研究者らによる確認を必要としており、我々のシステムではアノテーションデータを更新する機能も提供している。ここではPDB IDの対応付けのアノテーションを例にその手順を示す。

Annotation SystemメニューでPDB annotation workflowをクリックすると右のような画面が表示される。Workflowは多数の配列に対するアノテーションを効率的に進めるためのフレームワークで、画面左上のWorkFlowのところにあるように、3つのフェーズからなるユーザーインターフェースで構成されている。

まずはターゲットとなる配列群を探し出すRegion Search  
フェーズで、閲覧の時に使用したRegionSearchと同様の画面  
を用いて検索する。事前に処理してあるホモロジーの結果から  
アノテーションするのでここではRevHomologのタブをクリック  
して対象の配列を探す。

例としてBiosynthesis Enzymeのキーワードが入っているPDBエントリーにヒットした配列を探してみる。KeywordにBiosynthesis Enzymeと入力し、PDB HitのHitを選択、percent identityをLE (Lower Equal) を選択して60.0と入力してSearchをクリックする。

結果として1UC8と1UC9のエントリーが見つかる。これらをRegionListに登録するためにadd all regionsをクリックする。ターゲットとなる配列のIDはいずれも同じものだったため、重複が自動的に判別されてひとつのIDがRegionListにいれる。

これでアノテーションの準備ができた。検索ツールのすぐ下にあるnextボタンをクリックしてPDB annotationのフェーズへうつる。

**Workflow List**

- Region Search**
- PDB Annotations
- Region

**Region List**

0 entries:

[reg id](#)

[Delete all regions](#)

**Region Search**

**Simple** **BLAST** **Advance** **Revlookup** **File**

TtID Hi: HI < NonHi < Any < TtId  
 PDB Hi: HI < NonHi < Any < PDB  
 g Hi: HI < NonHi < Any < gi  
 percent identity: GE < LE < Any < 60.0 %  
 score: GE < LE < Any < 0.0  
 eval: GE < LE < Any < 0.0

when you done the searching and picking up regions go [next](#)

(show results by 10)

4 entries:

☐ PDB ☐ strand ☐ pos\_start ☐ pos\_end ☐ dna\_len ☐ ref\_id ☐ Option

|                     | reg_id          | hit_id          | percent_identity | score | eval |
|---------------------|-----------------|-----------------|------------------|-------|------|
| <a href="#">seq</a> | TtID8404f312000 | TtID8404f312000 | 95.0             | 498.8 | 0.0  |
| <a href="#">seq</a> | TtID8404f320000 | TtID8404f320000 | 95.0             | 498.8 | 0.0  |
| <a href="#">seq</a> | TtID8404f330000 | TtID8404f330000 | 94.29            | 495.0 | 0.0  |
| <a href="#">seq</a> | TtID8404f330000 | TtID8404f330000 | 94.29            | 499.4 | 0.0  |

[add all regions](#)

The screenshot shows the BLAST search interface with the 'Simple' tab selected. The search term is 'Biosynthesis Enzyme'. The results show a top hit with a 100% identity, 100% coverage, and an e-value of 0.0.

4 entries:

|     | PDB               | strand | pos_start | pos_end  | dna_len | ref_id   | Options |
|-----|-------------------|--------|-----------|----------|---------|----------|---------|
|     | reg_id            | hit_id | percent   | identity | score   | evaluate |         |
| add | TT-HB84orR1312300 | 1UCB   | 95.0      |          | 498.8   | 0.0      |         |
| add | TT-HB84orR1312300 | 1UCB   | 95.0      |          | 498.8   | 0.0      |         |
| add | TT-HB84orR1312300 | 1UC9   | 94.29     |          | 495.4   | 0.0      |         |
| add | TT-HB84orR1312300 | 1UC9   | 94.29     |          | 495.4   | 0.0      |         |

add all regions

Region List

1 entries:

| reg_id                      |
|-----------------------------|
| <del>TTHB84orR1312300</del> |

delete all regions

PDB annotationフェーズは右の画面のような構成になっている。左側は先ほどのRegion Searchフェーズと同じ構成でWorkFlowのところはPDB annotationフェーズのところに色がついて表示されている。右側はPDB IDの対応を入力するアノテーションインターフェースで、その下は閲覧の時と同じような配列情報を表示するツールになっている。右の画面の例ではHomologsタブをクリックしPDBタブをクリックしてPDBエントリーとのホモロジー情報を表示している。

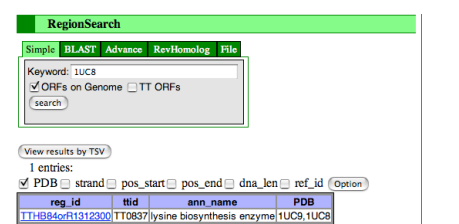
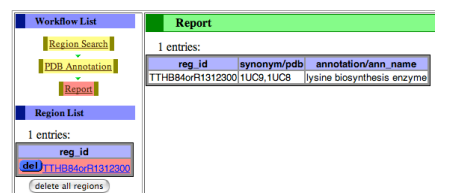
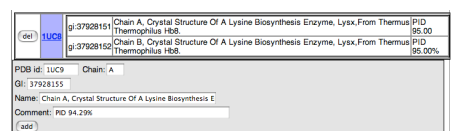
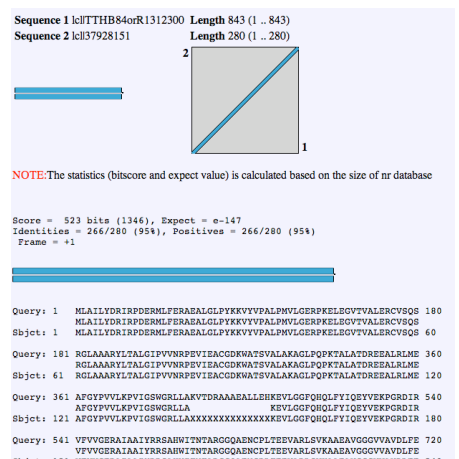
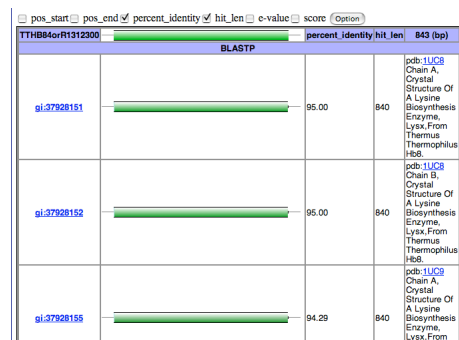
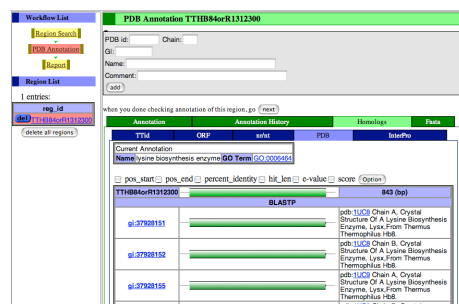
Homologsツール中ではターゲット配列とホモロジーが見られたエントリーのより細かい情報が見られるようになっているがデフォルトではいくつかの情報が表示されていない。例えばpercent identityとhit len（アライメントがとられた範囲の配列の長さ）を同時に見たい場合はホモロジー情報の上の各項目のチェックボックスにチェックをしてOptionをクリックすることでこれらの情報が表示される。この例ではORFの長さが843bpなのに対してhit lenが840なので1コドンだけ足りないことがわかるし、ホモロジーのバーの右端に少し黒い帯が見えていることから終端のコドンが外されているようにも見える。

アライメントを詳細に見たい場合はこのバーをクリックすることで詳細なアライメントを見ることができる。アライメントはBLAST2Sequenceを使ってリアルタイムに計算されるので必要であれば各種パラメーターを調整して再計算させることもできる。

ほかにもホモロジー情報のところのgi:のところはNCBIのEntrezにリンクしてあってこのエントリーの詳細を見ることができるし、pdb:のところはPDBのデータベースにリンクしてあってより詳細な情報をおいかけていくことができる。

このようにしてホモロジー情報を確認して当該のPDB IDがこのORFに付けられるべきであると判定できたらアノテーションインターフェースの空欄にPDBやChain名など必要事項を入力してaddをクリックすることでPDBアノテーションを追加することができる。またアノテーションインターフェースの上にすでに付けられたアノテーション情報が表示されているのでこれを見て確認することもできるし、PDBの横にあるDelボタンをクリックすることでこのアノテーションをORFから外すことができる。アノテーションがすんだらnextをクリックすることで次の配列へ移動するが、この例では配列は一つだったのでReportフェーズへうつり、今回のアノテーション結果をまとめて確認できる。

アノテーション情報は即座にデータベースへ反映されるため、この後すぐに閲覧の画面等でRegion Searchを使って先ほどのPDB IDを検索するとこの配列のIDが見つかる。



その他のツール

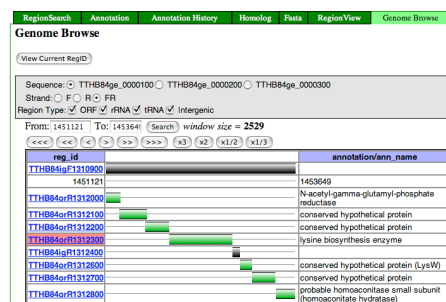
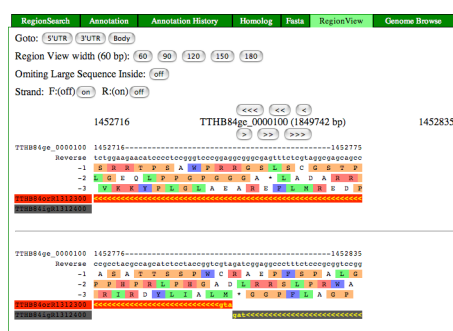
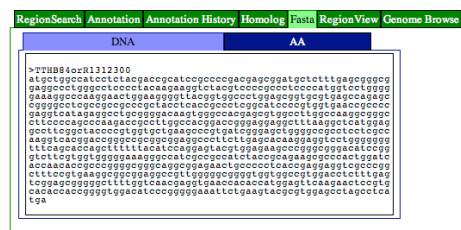
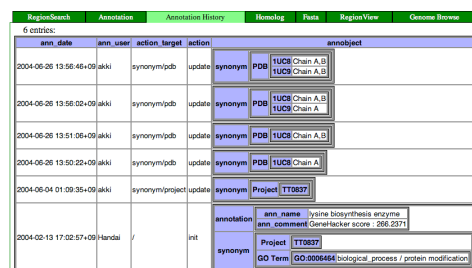
本システムでは他にも情報の閲覧やアノテーションを助けるツールが用意されている。

Annotation Historyではアノテーションが付けられていった経緯を見ることができるため、判定が難しいアノテーションではどのような意見が過去に出てアノテーションされたかを見ることができる。

FastaではNucleotide(DNA)とPeptide(AA)の配列をFasta形式で取得することができる。BLAST等のツールへここでコピーした配列をペーストして使用することができる。Peptideの配列はORFが設定されている読み方向にたいして3段階のフレームシフト全部を用意してある。

RegionViewではORFの配列の様子をみることができる。DNA配列とAA配列の両方がゲノム上のマップで表示されており、他の領域とどのように重なっているのかなど詳細にみることができる。また5'UTR、3'UTRのボタンを用意してあるのでORFの両末端付近の配列を素早く確認することもできる。さらに表示位置を移動したり、表示範囲を広げたりすることができるため、ORFのところだけでなくさらに上流や下流をずっと追っていくことができるようになっている。

Genome BrowseではView current RegIDをクリックすることでそのORFの近辺のゲノム情報を見ることができる。データベースに登録されているORFやORFの間のIntergenic領域がエントリーとして表示されそれらの配置がバーで表示されるため、各配列の大きさや位置関係が直感的に把握できるようになっている。また機能名をリストで見やすく表示してあるのでターゲットの遺伝子の前後にどのような機能のものがあるかを簡単に確認できるようになっている。このツールにも参照場所をスクロールさせるボタンやズームイン、ズームアウトがあるので自由に表示領域を変更することができる。



まとめ、今後

遠隔地から閲覧できるアノテーションデータベースという課題を満たすため我々はWeb技術を用いてアノテーションシステムを構築した。今回は基本的なツールと基礎的なデータ構造やシステムアーキテクチャの検討を行いながら実装をしていくことが目標であったが最低限のものは用意できたように思われる。しかし、“丸ごと一匹”を実現するためにはこれから先より多くの知見を効率的に蓄積し、利用できるよう環境整備をすすめていかななくてはならない。特に代謝系の情報やMS、GeneChipなどの発現データなど実験の結果などを集約していけるように速急に対応していく予定である。