

データベース統合に関わる基盤技術開発

「TogoWS」

詳細設計書

平成 26 年 4 月 24 日

大学共同利用機関法人 情報・システム研究機構
ライフサイエンス統合データベースセンター

目次

1. TogoWS 検索 API	4
1.1. 検索対象データベース一覧.....	4
1.2. キーワード検索.....	8
1.3. 検索結果の件数取得.....	9
1.4. 検索結果の形式指定.....	9
1.4.1. JSON 形式での結果取得.....	9
1.4.2. HTML 形式での結果取得.....	10
2. TogoWS 取得 API	11
2.1. 取得対象データベース一覧.....	11
2.2. エントリ取得.....	12
2.3. エントリの部分取得.....	12
2.4. 配列の部分取得.....	13
2.5. エントリの形式変換.....	15
2.5.1. JSON 形式での結果取得.....	16
2.5.2. FASTA 形式での取得.....	18
2.5.3. GFF 形式での取得.....	18
2.5.4. XML 形式での取得.....	19
2.5.5. Turtle 形式での取得.....	19
3. TogoWS 変換 API	21
3.1. 変換対象データ形式一覧.....	21
3.2. データ形式変換.....	22
3.3. RDF データへの形式変換.....	27
3.3.1. Turtle から RDF/XML への変換.....	28
3.3.2. RDF/XML から Turtle への変換.....	28
3.3.3. CSV 形式から RDF への変換.....	28
3.3.4. GenBank 形式から RDF への変換.....	29
3.3.5. EMBL 形式から RDF への変換.....	29
3.3.6. DDBJ 形式から RDF への変換.....	29
3.3.7. HMMER3 table から RDF への変換.....	29
4. TogoWS UCSC API	30
4.1. UCSC の対応データベース.....	30
4.1.1. データベースの一覧.....	31
4.1.2. テーブルの一覧.....	31

4.1.3. カラムの一覧	32
4.2. キーワード検索	34
4.3. 領域指定検索	37
4.3.1. 領域検索の境界指定	38
4.4. CDS, exon, intron の検索	40
4.5. bigWig ファイルの検索	41
4.6. bigBed ファイルの検索	44
4.7. 2bit ファイルからの塩基配列の取得	47

1. TogoWS 検索 API

検索APIでは対象となるデータベースにキーワード検索を行い該当するエントリのIDを取得するサービスを提供する。

1.1. 検索対象データベース一覧

対象となるデータベースの一覧は下記の REST API で取得できる。

<http://togows.org/search>

結果として、下記のデータベース一覧が取得される。第1カラムがデータベースの正式名称で第2カラムが省略名である。

pdbj-pdb	pdb
kegg-compound	compound
kegg-drug	drug
kegg-enzyme	enzyme
kegg-genes	genes
kegg-glycan	glycan
kegg-orthology	orthology
kegg-reaction	reaction
kegg-module	module
kegg-pathway	pathway
ncbi-pubmed	pubmed
ncbi-protein	protein
ncbi-nuccore	nuccore
ncbi-nucleotide	nucleotide
ncbi-nucgss	nucgss
ncbi-nucest	nucest
ncbi-structure	
ncbi-genome	
ncbi-assembly	
ncbi-gcassembly	
ncbi-genomeprj	
ncbi-bioproject	
ncbi-biosample	
ncbi-biosystems	

ncbi-blastdbinfo
ncbi-books
ncbi-cdd
ncbi-clone
ncbi-gap
ncbi-gapplus
ncbi-dbvar
ncbi-epigenomics
ncbi-gene gene
ncbi-gds
ncbi-geoprofiles
ncbi-homologene homologene
ncbi-journals
ncbi-medgen
ncbi-mesh mesh
ncbi-ncbisearch
ncbi-nlmcatalog
ncbi-omia
ncbi-omim omim
ncbi-pmc
ncbi-popset
ncbi-probe
ncbi-proteinclusters
ncbi-pcassay
ncbi-pccompound
ncbi-pcsubstance
ncbi-pubmedhealth
ncbi-seqannot
ncbi-snp snp
ncbi-sra
ncbi-taxonomy
ncbi-toolkit
ncbi-toolkitall
ncbi-toolkitbook
ncbi-unigene
ncbi-unists

ncbi-gencoll
ebi-embl embl
ebi-arrayexpress-repository
ebi-atlas-experiments
ebi-atlas-genes
ebi-biomodels
ebi-chebi
ebi-chembl-activity
ebi-chembl-assay
ebi-chembl-target
ebi-dgva
ebi-efo
ebi-ega
ebi-emblcds
ebi-emblnew_con
ebi-emblnew_standard
ebi-emblnew_tsa
ebi-emblnew_wgs
ebi-emblrelease_con
ebi-emblrelease_standard
ebi-emblrelease_tsa
ebi-emblrelease_wgs
ebi-ensemblGenomes_gene
ebi-ensembl_gene
ebi-epo
ebi-genome_assembly
ebi-go
ebi-gpcrdb
ebi-hgnc
ebi-intact-experiments
ebi-intact-interactions
ebi-intact-interactors
ebi-intenz
ebi-interpro
ebi-jpo
ebi-kipo

ebi-lrg
ebi-medline
ebi-merops_clan
ebi-merops_family
ebi-merops_id
ebi-nrn11
ebi-nrn12
ebi-nrp11
ebi-nrp12
ebi-omim
ebi-patentFamilies
ebi-patentdb
ebi-pdbe
ebi-pdbechem
ebi-pride
ebi-project
ebi-reactome
ebi-resid
ebi-rhea
ebi-sbo
ebi-sra-analysis
ebi-sra-experiment
ebi-sra-run
ebi-sra-sample
ebi-sra-study
ebi-sra-submission
ebi-taxonomy
ebi-uniparc uniparc
ebi-uniprot uniprot
ebi-uniref100 uniref100
ebi-uniref50 uniref50
ebi-uniref90 uniref90
ebi-uspto
ebi-wgs_masters

1.2. キーワード検索

キーワード検索はデータベース名とキーワードを指定して下記の REST API で行う。

<http://togows.org/search/データベース名/キーワード/limit,offset>

ここで **limit** は検索開始件数、**offset** は一度に取得する件数である。

具体的に UniProt データベースからキーワード **lung cancer** で該当するエントリを最初の 10 件取得するには下記の API を利用する。

<http://togows.org/search/uniprot/lung%20cancer/1,10>

結果として下記の ID 一覧が得られる。

DLEC1_HUMAN
KKLC1_MACFA
DLEC1_RAT
DLEC1_MOUSE
KKLC1_HUMAN
LAS2_MOUSE
BLCAP_HUMAN
CT55_HUMAN
CT455_HUMAN
SDCG8_HUMAN

さらに、続きの 5 件を取得するには下記の REST API を使用する。

<http://togows.org/search/ebi-uniprot/lung%20cancer/11,5>

結果として下記の 5 件のエントリ ID が取得される。

HIC2_MOUSE
RCAS1_HUMAN
MACC1_HUMAN
CASC1_MOUSE
CASC1_HUMAN

1.3. 検索結果の件数取得

検索結果の件数は下記の REST API で取得できる。

```
http://togows.org/search/データベース名/キーワード/count
```

具体的に UniProt データベースからキーワード **lung cancer** で該当するエントリの数を取
得するには下記の API を利用する。

```
http://togows.org/search/uniprot/lung%20cancer/count
```

結果として下記のように該当するエントリの件数が得られる。

```
1045
```

1.4. 検索結果の形式指定

検索結果は上記のようにテキスト形式で取得するほか、JavaScript などを利用しやすい
JSON 形式や、TogoWS の取得 API に対応しているデータベースでは、ブラウザから利用
しやすい HTML 形式での取得も可能である。

検索対象データベースが対応する形式の一覧は下記の REST API で得られる。

```
http://togows.org/search/データベース名?formats
```

例として UniProt データベースで対応形式を取得する場合は下記のようになる。

```
http://togows.org/search/uniprot?formats
```

結果として下記が得られ、JSON 形式と HTML 形式に対応していることが分かる。

```
json  
html
```

1.4.1. JSON 形式での結果取得

JSON 形式で結果を取得するには、上記の REST API に拡張子 `.json` を追記する。

```
http://togows.org/search/uniprot/lung%20cancer/1,10.json
```

結果として下記のような JSON 形式のデータが取得される。

```
["DLEC1_HUMAN", "KKLC1_MACFA", "DLEC1_RAT", "DLEC1_MOUSE", "KKLC1_HUMAN", "LAS2_M  
OUSE", "BLCAP_HUMAN", "CT55_HUMAN", "CT455_HUMAN", "SDCG8_HUMAN"]
```

1.4.2. HTML 形式での結果取得

HTML 形式で結果を取得するには、上記の REST API に拡張子 `.html` を追記する。

<http://togows.org/search/uniprot/lung%20cancer/1,10.html>

結果として図のように HTML 形式で結果が得られ、とくにウェブブラウザからアクセスする際には、検索結果からクリックするだけで該当するエントリを取得することができる。

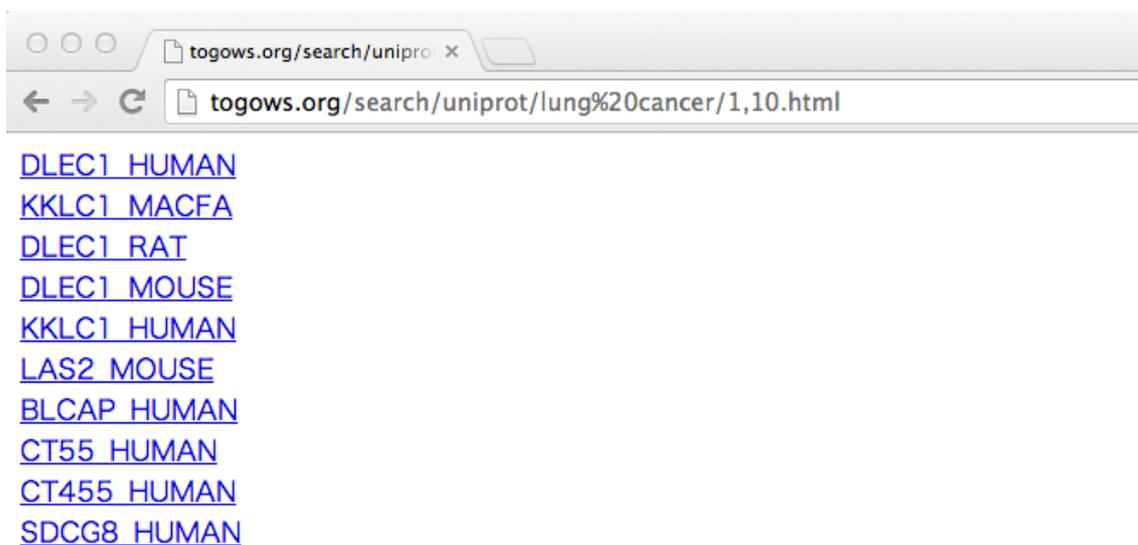


図 : TogoWS の検索 API において出力結果を HTML 形式で取得する例

2. TogoWS 取得 API

取得 API では対象となるデータベースのエントリを取得し、さらに形式変換や部分情報の抽出などを行なうサービスを提供する。

2.1. 取得対象データベース一覧

対象となるデータベースの一覧は下記の REST API で取得できる。

<http://togows.org/entry>

結果として、下記のデータベース一覧が取得される。第 1 カラムがデータベースの正式名称で第 2 カラムが省略名である。

ncbi-nuccore	nuccore
ncbi-nucest	nucest
ncbi-nucgss	nucgss
ncbi-nucleotide	nucleotide
ncbi-protein	protein
ncbi-gene	gene
ncbi-homologene	homologene
ncbi-snp	snp
ncbi-mesh	mesh
ncbi-pubmed	pubmed
ebi-embl	embl
ebi-uniprot	uniprot
ebi-uniparc	uniparc
ebi-uniref100	uniref100
ebi-uniref90	uniref90
ebi-uniref50	uniref50
ddbj-ddbj	ddbj
ddbj-dad	dad
pdbj-pdb	pdb
kegg-compound	compound
kegg-drug	drug
kegg-enzyme	enzyme
kegg-genes	genes
kegg-glycan	glycan

```
kegg-orthology      orthology
kegg-reaction       reaction
kegg-module         module
kegg-pathway        pathway
```

2.2. エントリ取得

エントリの取得はデータベース名とエントリの ID を指定して下記の REST API で行なう。

<http://togows.org/entry/データベース名/エントリ ID>

エントリ ID はカンマで区切って複数指定することも可能である。

<http://togows.org/entry/データベース名/エントリ ID1,エントリ ID2>

具体的に NCBI 塩基配列データベース nucleotide からエントリ NC_001138 を取得するには下記の REST API を利用する。

http://togows.org/entry/nucleotide/NC_001138

2.3. エントリの部分取得

さらに、エントリの一部を抽出するには下記の REST API を使用する。

<http://togows.org/entry/データベース名/エントリ ID/部分名>

ここで指定可能な部分名はデータベースごとに異なるが、下記の REST API で取得することができる。

<http://togows.org/entry/データベース名?fields>

NCBI の nucleotide データベースの場合には、

<http://togows.org/entry/nucleotide?fields>

により下記の部分名が指定可能であることが分かる。

```
entry_id
```

length
strand
moltype
linearity
division
date
definition
accession
accessions
version
versions
acc_version
gi
keywords
organism
common_name
taxonomy
comment
seq
references
features

具体的にエントリ NC_001138 の **definition** を取得する場合は、

http://togows.org/entry/nucleotide/NC_001138/definition

により

Saccharomyces cerevisiae S288c chromosome VI, complete sequence.

が得られ、出芽酵母の染色体 6 番の配列のエントリであることが分かる。

2.4. 配列の部分取得

配列のエントリの場合、**seq** を指定することでエントリ全体の配列が取得出来るだけでなく、**INSDC** の座標フォーマットを指定することによりスプライシングなど複雑な部分配列の取得も可能となっている。

具体的には、上記の例のエントリ NC_001138 では、

http://togows.org/entry/ncbi-nucleotide/NC_001138/seq

により染色体 6 番の全体配列 270161 塩基対が取得できるが、

```
LOCUS      NC_001138          270161 bp   DNA    linear   PLN 06-SEP-2013
DEFINITION Saccharomyces cerevisiae S288c chromosome VI, complete sequence.
(省略)
FEATURES             Location/Qualifiers
(省略)
     mRNA             complement(join(53260..54377,54687..54696))
                       /gene="ACT1"
                       /locus_tag="YFL039C"
                       /gene_synonym="ABY1; END7"
                       /product="actin"
```

上記に示されるようにアクチンの遺伝子 ACT1 の配列を取得するには相補鎖にある 2 つの部分配列をつなぎ合わせる必要がある。TogoWS では

[http://togows.org/entry/ncbi-nucleotide/NC_001138/seq/complement\(join\(53260..54377,54687..54696\)\)](http://togows.org/entry/ncbi-nucleotide/NC_001138/seq/complement(join(53260..54377,54687..54696)))

のように INSDC の座標フォーマットを指定することで、下記のように ACT1 遺伝子の全長 1128 塩基対を取得することができる。

```
atggattctgaggttgctgcttttggttattgataacggttctgggtatgtgtaaagccggttttgccggtgacgacgc
tcctcgtgctgtcttcccattctatcgctcggttagaccaagacaccaaggtatcatgggtcggtatgggtcaaaaagact
cctacggtgggtgatgaagctcaatccaagagaggtatcttgactttacgttaccaattgaacacggtattgtcacc
aactgggacgatatggaaaagatctggcatcataccttctacaacgaattgagagttgccccagaagaacacctgt
tcttttgactgaagctccaatgaaccctaaatcaaacagagaaaagatgactcaaattatgtttgaaactttcaacg
ttccagccttctacgcttccatccaagccggttttgctccttgactcttccggtagaactactgggtattgtttggat
tccggtgatgggtgttactcacgctcgttccaatctacgctgggtttctctctacctcacgccatcttgagaatcgattt
ggccggtagagatttgactgactacttgatgaagatcttgagtgaacgtgggttactcttctccaccactgctgaaa
gagaaattgtccgtgacatcaaggaaaaactatgttacgctcgccttgacttcgaacaagaaatgcaaaccgctgct
```

```
caatcttcttcaattgaaaaatcctacgaacttccagatgggtcaagtcactattggtaacgaaagattcagagc
cccagaagctttgttccatccttctgttttgggtttggaatctgccggtattgaccaaactacttacaactccatca
tgaagtgtgatgtcgatgtccgtaaggaattatacggtaacatcgttatgtccggtggtagaccaccatgttcccaggt
attgccgaaagaatgcaaaggaaatcacgctttggctccatcttccatgaagggtcaagatcattgctcctccaga
aagaaagtactccgctcgattgggtggttctatcttggcttctttgactaccttccaacaaatgtggatctcaaac
aagaatacgacgaaagtgggtccatctatcgttcaccacaagtgtttctaa
```

2.5. エントリの形式変換

エントリは上記のようにテキスト形式で取得する他、データベースごとに **JSON** 形式などいくつかの形式に変換して取得することも可能である。

取得対象データベースが対応する形式の一覧は下記の **REST API** で得られる。

<http://togows.org/entry/データベース名?formats>

例として NCBI の **nucleotide** データベースで対応形式を取得する場合は下記のようになる。

<http://togows.org/entry/nucleotide?formats>

結果として下記が得られ、**JSON** 形式、**FASTA** 形式、**GFF** 形式、**XML** 形式、**GenBank** 形式(TogoWS では **INSDC** の **CON division** などは **gbwithparts** オプションを指定して全配列を結合して取得しているが、オリジナルの **gb** 形式で取得したい場合に利用)、**Turtle** 形式(セマンティック・ウェブの **RDF** フォーマット)に対応していることが分かる。

```
fasta
gff
json
xml
gb
ttl
```

これにより、下記の **REST API** により対象エントリを形式変換して取得できる。

<http://togows.org/entry/データベース名/エントリ ID.対応形式>

また、これらの形式変換はエントリの部分取得と組み合わせることも可能である。

<http://togows.org/entry/データベース名/エン트리 ID/部分名.対応形式>

2.5.1. JSON 形式での結果取得

JSON 形式で上記のエントリを取得する場合、REST API に拡張子 `.json` を追記する。

http://togows.org/entry/ncbi-nucleotide/NC_001138.json

さらにエントリの一部分だけを対象に形式変換して取得することも可能で、

http://togows.org/entry/ncbi-nucleotide/NC_001138/references.json

の例ではエントリのリファレンス部分だけを JSON 形式で取得する。

```
[
  [
    {
      "authors": [ "Goffeau, A.", "Barrell, B.G.", "Bussey, H.", "Davis, R.W.",
        "Dujon, B.", "Feldmann, H.", "Galibert, F.", "Hoheisel, J.D.", "Jacq, C.",
        "Johnston, M.", "Louis, E.J.", "Mewes, H.W.", "Murakami, Y.", "Philippsen, P.",
        "Tettelin, H.", "Oliver, S.G." ]
      ,
      "title": "Life with 6000 genes",
      "journal": "Science 274 (5287), 546 (1996)",
      "volume": "",
      "issue": "",
      "pages": "",
      "year": "",
      "pubmed": "8849441",
      "medline": "",
      "doi": null,
      "abstract": "",
      "url": null,
      "mesh": [ ],
      "embl_gb_record_number": 1,
      "sequence_position": "1-270161",
```

```

    "comments": null,
    "affiliations": [ ]
  },
  {
    "authors": [ "Murakami, Y.", "Naitou, M.", "Hagiwara, H.", "Shibata, T.",
"Ozawa, M.", "Sasanuma, S.", "Sasanuma, M.", "Tsuchiya, Y.", "Soeda, E.",
"Yokoyama, K. et al." ],
    "title": "Analysis of the nucleotide sequence of chromosome VI from
Saccharomyces cerevisiae",
    "journal": "Nat. Genet.",
    "volume": "10",
    "issue": "3",
    "pages": "261-268",
    "year": "1995",
    "pubmed": "7670463",
    "medline": "",
    "doi": null,
    "abstract": "",
    "url": null,
    "mesh": [ ],
    "embl_gb_record_number": 2,
    "sequence_position": "1-270161",
    "comments": null,
    "affiliations": [ ]
  },
  {
    (省略)
  }
]

```

また複数エントリについて説明や塩基配列長を JSON 形式で取得する場合は下記のようになる。

```

http://togows.org/entry/ncbi-nucleotide/NC_001138,NC_001139/definition.json
["Saccharomyces cerevisiae S288c chromosome VI, complete sequence.",
"Saccharomyces cerevisiae S288c chromosome VII, complete sequence."]

```

http://togows.org/entry/ncbi-nucleotide/NC_001138,NC_001139/length.json
[270161, 1090940]

2.5.2. FASTA 形式での取得

FASTA 形式で上記のエントリを取得する場合、REST API に拡張子 `.fasta` を追記する。

http://togows.org/entry/ncbi-nucleotide/NC_001138.fasta

結果は下記のようになる。

```
>gi|330443543|ref|NC_001138.5| Saccharomyces cerevisiae S288c chromosome VI,
complete sequence
GATCTCGCAAGTGCATTCCTAGACTTAATTCATATCTGCTCCTCAACTGTCGATGATGCCTGCTAAACTG
CAGCTTGACGTACTGCGGACCCTGCAGTCCAGCGCTCGTCATGGAACGCAAACGCTGAAAACTCCAAC
TTCTCGAGCGCTTCCACAAAGACCGTATCGTCTTTTGCCTCCCATTCTTCCCGGCACTTTTTCTCGTCCC
AGTTCAAAAAGTACTGCAGCACCTCTGTCTTCGATTACGCAAGTTGCTCCATACTTTATAATAACAAC
(省略)
```

2.5.3. GFF 形式での取得

GFF 形式で上記のエントリを取得する場合、REST API に拡張子 `.gff` を追記する。

http://togows.org/entry/ncbi-nucleotide/NC_001138.gff

結果は下記のようになる。

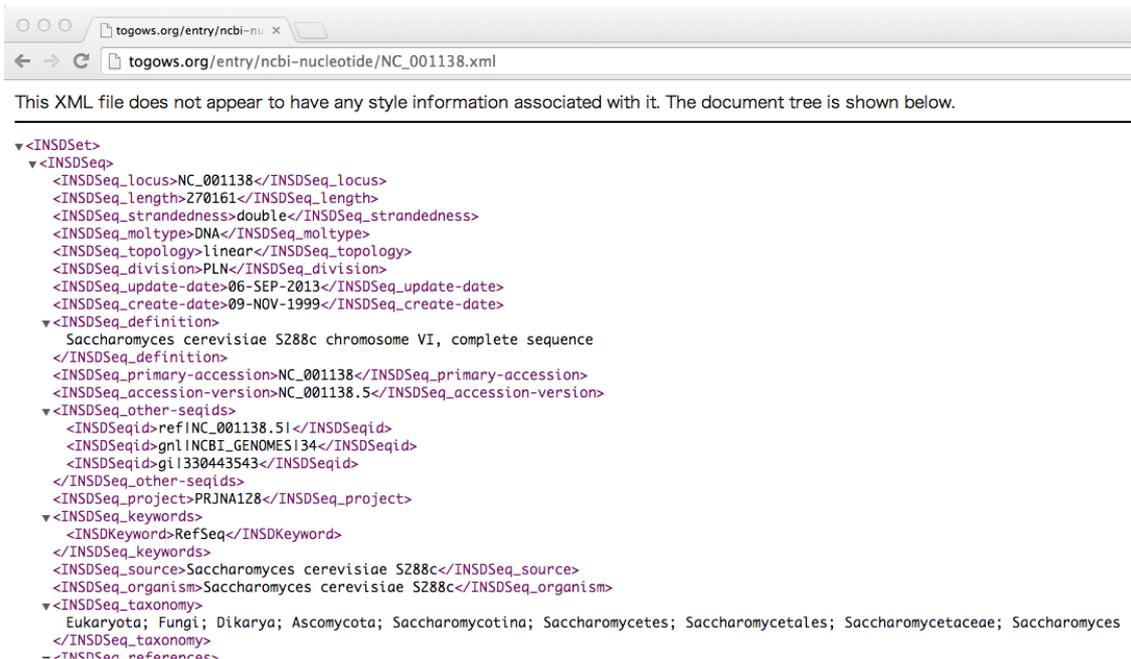
```
##gff-version 3
NC_001138 Genbank region 1 270161 . . . ID=NC_001138;Note=Saccharomyces%20cerevisiae%20S288c%
NC_001138 Genbank region 1 270161 . + . ID=Saccharomyces%20cerevisiae%20S288c;db_xref=taxon%3A5
NC_001138 Genbank telomere 1 5530 . - . ID=TEL06L%3B%20Telomeric%20region%20on%20the%20left%2
NC_001138 Genbank repeat_region 1 4684 . - . ID=Y%27%20element;note=TEL06L-YP%3B%20Telomeric%20sh
NC_001138 Genbank gene 53 535 . + . ID=YFL068W;db_xref=GeneID%3A850476
NC_001138 Genbank mRNA 53 535 . + . Parent=YFL068W;ID=YFL068W.t01;product=hypothetical%20prote
NC_001138 Genbank CDS 53 535 . + . Parent=YFL068W.t01;db_xref=GI%3A14318451,SGD%3AS00000
NC_001138 Genbank gene 836 1363 . + . ID=YFL067W;db_xref=GeneID%3A850477
NC_001138 Genbank mRNA 836 1363 . + . Parent=YFL067W;ID=YFL067W.t01;transcript_id=NM_001179900.
NC_001138 Genbank CDS 836 1363 . + . Parent=YFL067W.t01;db_xref=GI%3A14318452,SGD%3AS00000
NC_001138 Genbank gene 1437 2615 . - . ID=YFL066C;db_xref=GeneID%3A850478
NC_001138 Genbank mRNA 1437 2615 . - . Parent=YFL066C;ID=YFL066C.t01;db_xref=GI%3A296144367,Ge
NC_001138 Genbank CDS 1437 2615 . - . Parent=YFL066C.t01;db_xref=GI%3A14318453,SGD%3AS000001
NC_001138 Genbank gene 3030 3338 . - . ID=YFL065C;db_xref=GeneID%3A850479
```

2.5.4. XML 形式での取得

XML 形式で上記のエントリを取得する場合、REST API に拡張子 `.xml` を追記する。

`http://togows.org/entry/ncbi-nucleotide/NC_001138.xml`

結果は下記のようになる。



2.5.5. Turtle 形式での取得

Turtle 形式で上記のエントリを取得する場合、REST API に拡張子 `.ttl` を追記する。

`http://togows.org/entry/ncbi-nucleotide/NC_001138.ttl`

結果は下記のようになる。

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix obo: <http://purl.obolibrary.org/obo/> .
@prefix faldo: <http://biohackathon.org/resource/faldo#> .
@prefix insdc: <http://ddbj.nig.ac.jp/ontologies/sequence/> .
```

```
<http://identifiers.org/bioproject/PRJNA128>  
  a <http://identifiers.org/bioproject> .
```

```
<http://identifiers.org/ncbigene/1466401>  
  a <http://identifiers.org/ncbigene> ;  
  rdfs:label "1466401" .
```

```
<http://identifiers.org/ncbigene/1466402>  
  a <http://identifiers.org/ncbigene> ;  
  rdfs:label "1466402" .
```

(省略)

3. TogoWS 変換 API

変換 API ではユーザが HTTP POST したデータを別のデータ形式に変換するサービスを提供する。

3.1. 変換対象データ形式一覧

変換可能なデータ形式の組み合わせは下記の REST API で取得できる。

<http://togows.org/convert>

結果として、下記の組み合わせ一覧が取得される。

```
genbank.fasta
genbank.embl
genbank.gff
genbank.ntriples
genbank.n3
genbank.rdfxml
genbank.ttl
embl.fasta
embl.genbank
embl.ntriples
embl.n3
embl.rdfxml
embl.ttl
ddbj.ntriples
ddbj.n3
ddbj.rdfxml
ddbj.ttl
uniprot.fasta
uniprot.gff
blast.gff
blasttable.gff
blastxml.gff
megablast.gff
fasta.gff
psl.gff
```

```
sim4.gff
hmmer.gff
hmmer3tbl.rdfxml
exonerate.gff
wise.gff
rdfxml.ttl
ttl.rdfxml
csv.rdfxml
csv.ttl
gff.rdfxml
gff.ttl
gvf.rdfxml
gvf.ttl
```

3.2. データ形式変換

データの形式変換には入力データ形式と出力データ形式を指定して下記の REST API を使用する。なお、HTTP GET ではなく、HTTP POST を使用しファイルをサーバに送る。結果として形式変換後のデータが返される。

```
http://togows.org/convert/入力データ形式.出力データ形式
```

具体的に GenBank 形式のデータを EMBL 形式に変換するには、まず GenBank 形式のデータをファイルに用意し、

```
% curl -O http://togows.org/entry/nucleotide/J00231
```

このファイルを下記の REST API に POST する。

```
% curl http://togows.org/convert/genbank.embl --data-binary @J00231
```

結果、下記の形式変換が行われる。

入力データ :

```
LOCUS      HUMIGHAF                1089 bp   mRNA   linear   PRI 09-NOV-1994
DEFINITION Human Ig gamma3 heavy chain disease OMM protein mRNA.
ACCESSION  J00231
```

VERSION J00231.1 GI:185041
 KEYWORDS C-region; V-region; gamma heavy chain disease protein; gamma3 heavy chain disease protein; heavy chain disease; hinge exon; immunoglobulin gamma-chain; immunoglobulin heavy chain; secreted immunoglobulin.
 SOURCE Homo sapiens (human)
 ORGANISM Homo sapiens
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
 REFERENCE 1 (bases 1 to 1089)
 AUTHORS Alexander,A., Steinmetz,M., Barritault,D., Frangione,B., Franklin,E.C., Hood,L. and Buxbaum,J.N.
 TITLE gamma Heavy chain disease in man: cDNA sequence supports partial gene deletion model
 JOURNAL Proc. Natl. Acad. Sci. U.S.A. 79 (10), 3260-3264 (1982)
 PUBMED 6808505
 COMMENT Original source text: Human (patient OMM) peripheral blood cells, cDNA to mRNA.
 The protein isolated from patient OMM is a gamma heavy chain disease (HCD) protein. It has a large 5' internal deletion consisting of most of the variable region and the entire ch1 domain. [1] suggests that the protein abnormality is from a partial gene deletion rather than from defective splicing.
 FEATURES Location/Qualifiers
 source 1..1089
 /organism="Homo sapiens"
 /mol_type="mRNA"
 /db_xref="taxon:9606"
 /map="14q32.33"
 gene 1..1089
 /gene="IGHG3"
 mRNA <1..1089
 /gene="IGHG3"
 /product="gamma3 mRNA"
 CDS 23..964

```

/gene="IGHG3"
/note="OMM protein (Ig gamma3) heavy chain"
/codon_start=1
/protein_id="AAA52805.1"
/db_xref="GI:567112"
/db_xref="GDB:G00-119-339"
/translation="MKXLWFFLLLVAAPRWVLSQVHLQESGPGLGKPKPELKTPLGDTT
HTCPRCPEPKSCDTPPPCPRCPEPKSCDTPPPCPRCPEPKSCDTPPPCPXCPAPPELLG
GPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPXVQFKWYVDGVEVHNAKTKLREE
QYNSTFRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTIKAKGQPXXXXXXXXXX
XXEEMTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYNTTPMLDSDGSFFLYSKL
TVDKSRWQQGNIFSCSVMEALHNRYTQKSLSLSPGK"
sig_peptide    26..79
/gene="IGHG3"
/note="OMM protein signal peptide"
mat_peptide    80..961
/gene="IGHG3"
/product="OMM protein mature peptide"

```

ORIGIN

```

1 cctggacctc ctgtgcaaga acatgaaaca nctgtggttc ttccttctcc tgggtggcagc
61 tcccagatgg gtccctgtccc aggtgcacct gcaggagtcg ggcccaggac tggggaagcc
121 tccagagctc aaaacccccac ttggtgacac aactcacaca tgcccacggt gccagagcc
181 caaatcttgt gacacacctc ccccggtgcc acggtgccc gagcccaaat cttgtgacac
241 acctcccca tgcccacggt gccagagcc caaatcttgt gacacacctc ccccggtgcc
301 nngtgccca gcacctgaac tcttgggagg accgtcagtc ttcctcttcc ccccaaaacc
361 caaggatacc cttatgattt cccggacccc tgaggtcacg tgcgtggttg tggacgtgag
421 ccacgaagac ccnnngtcc agttcaagtg gtacgtggac ggcgtggagg tgcataatgc
481 caagacaaag ctgcgggagg agcagtacaa cagcacgttc cgtgtggtca gcgtcctcac
541 cgtcctgcac caggactggc tgaacggcaa ggagtacaag tgcaaggtct ccaacaaagc
601 cctcccagcc cccatcgaga aaaccatctc caaagccaaa ggacagccn nnnnnnnnnn
661 nnnnnnnnnn nnnnnnnnnn nnnnngagga gatgaccaag aaccaagtca gcctgacctg
721 cctggtcaaa ggcttctacc ccagcgacat cgccgtggag tgggagagca atgggcagcc
781 ggagaacaac tacaacacca cgctcccat gctggactcc gacggctcct tcttctcta
841 cagcaagctc accgtggaca agagcaggtg gcagcagggg aacatcttct catgctccgt
901 gatgcatgag gctctgcaca accgctacac gcagaagagc ctctccctgt ctccgggtaa
961 atgagtgcc a tggccggcaa gccccgctc cccgggctct cggggtcgag cgaggatgct

```

1021 tggcacgtac cccgtgtaca tacttcccag gcacccagca tggaaataaa gcacccagcg
1081 ctgccctgg

//

出力データ :

ID J00231; SV 1; linear; mRNA; STD; UNC; 1089 BP.
XX
AC J00231;
XX
DT 09-NOV-1994
XX
DE Human Ig gamma3 heavy chain disease OMM protein mRNA.
XX
KW C-region; V-region; gamma heavy chain disease protein; gamma3 heavy; chain
KW disease protein; heavy chain disease; hinge exon; immunoglobulin
KW gamma-chain; immunoglobulin heavy chain; secreted; immunoglobulin
XX
OS Homo sapiens (human)
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC Homo.
XX
RN [1]
RP 1-1089
RX PUBMED; 6808505.
RA Alexander,A., Steinmetz,M., Barritault,D., Frangione,B., Franklin,E.C.,
RA Hood,L. and Buxbaum,J.N.;
RT gamma Heavy chain disease in man: cDNA sequence supports partial gene
RT deletion model;
RL Proc. Natl. Acad. Sci. U.S.A. 79 (10), 3260-3264 (1982)
XX
CC Original source text: Human (patient OMM) peripheral blood cells, cDNA to
CC mRNA. The protein isolated from patient OMM is a gamma heavy chain disease
CC (HCD) protein. It has a large 5' internal deletion consisting of most of
CC the variable region and the entire ch1 domain. [1] suggests that the
CC protein abnormality is from a partial gene deletion rather than from

```

CC    defective splicing.
XX
FH    Key          Location/Qualifiers
FH
FT    source       1..1089
FT                /db_xref="taxon:9606"
FT                /mol_type="mRNA"
FT                /map="14q32.33"
FT                /organism="Homo sapiens"
FT    gene         1..1089
FT                /gene="IGHG3"
FT    mRNA         <1..1089
FT                /gene="IGHG3"
FT                /product="gamma3 mRNA"
FT    CDS          23..964
FT                /translation="MKXLWFFLLLVAAAPRWVLSQVHLQESGPGGLGKPPPELKTPLGDTTH
FT                TPCRCPEPKSCDTPPPCPRCPEPKSCDTPPPCPRCPEPKSCDTPPPCPXCPAPELLGGP
FT                SVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPXVQFKWYVDGVEVHNAKTKLREEQYN
FT                STFRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQXXXXXXXXXXXXXE
FT                EMTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYNTTPPMLDSGDSFFLYSKLTVDKS
FT                RWQQGNIFSCSVMHREALHNRYSYQKSLSLSPGK"
FT                /codon_start=1
FT                /db_xref="GI:567112"
FT                /db_xref="GDB:G00-119-339"
FT                /protein_id="AAA52805.1"
FT                /note="OMM protein (Ig gamma3) heavy chain"
FT                /gene="IGHG3"
FT    sig_peptide  26..79
FT                /gene="IGHG3"
FT                /note="OMM protein signal peptide"
FT    mat_peptide  80..961
FT                /product="OMM protein mature peptide"
FT                /gene="IGHG3"
XX
SQ    Sequence 1089 BP; 240 A; 358 C; 271 G; 176 T; 44 other;
      cctggacctc ctgtgcaaga acatgaaaca nctgtggttc ttccttctcc tgggtggcgc 60

```

```

tcccagatgg gtcctgtccc aggtgcacct gcaggagtcg ggcccaggac tggggaagcc      120
tccagagctc aaaaccccac ttggtgacac aactcacaca tgcccacggt gcccagagcc      180
caaatcttgt gacacacctc ccccgtgccc acggtgcccc gagcccaaat cttgtgacac      240
acctcccca tgcccacggt gcccagagcc caaatcttgt gacacacctc ccccgtgccc      300
nngtgcccc gcacctgaac tcttgggagg accgtcagtc ttctcttcc ccccaaaacc      360
caaggatacc cttatgattt cccggacccc tgaggtcacg tgcgtggtgg tggacgtgag      420
ccacgaagac cennnngtcc agttcaagtg gtacgtggac ggcgtggagg tgcataatgc      480
caagacaaaag ctgcgggagg agcagtacaa cagcacgttc cgtgtggtca gcgtcctcac      540
cgtcctgcac caggactggc tgaacggcaa ggagtacaag tgcaaggctc ccaacaaagc      600
cctcccagcc cccatcgaga aaaccatctc caaagccaaa ggacagcccn nnnnnnnnnn      660
nnnnnnnnnn nnnnnnnnnn nnnnngagga gatgaccaag aaccaagtca gcctgacctg      720
cctggtcaaa ggcttctacc ccagcgacat cgccgtggag tgggagagca atgggcagcc      780
ggagaacaac tacaacacca cgcctcccat gctggactcc gacggctcct tcttctctta      840
cagcaagctc accgtggaca agagcagggtg gcagcagggg aacatcttct catgctccgt      900
gatgcatgag gctctgcaca accgctacac gcagaagagc ctctccctgt ctccgggtaa      960
atgagtgcc a tggccggcaa gccccgctc cccgggtct cggggtcgcg cgaggatgct     1020
tggcacgtac cccgtgtaca tacttcccag gcaccagca tggaaataaa gcaccagcg     1080
ctgcccctgg                                     1089

```

//

3.3. RDF データへの形式変換

生命科学で広く使われるデータ形式間の変換だけでなく、セマンティック・ウェブでのデータ統合のため、TogoWS では RDF へのデータ変換機能を提供している。RDF に変換できるファイル形式は以下の通り。

入力のファイル形式	出力のファイル形式
Turtle	RDF/XML
RDF/XML	Turtle
CSV	RDF/XML, Turtle
GenBank	RDF/XML, N-Triples, N3, Turtle
EMBL	RDF/XML, N-Triples, N3, Turtle
DDBJ	RDF/XML, N-Triples, N3, Turtle
HMMER3 table	RDF/XML

3.3.1. Turtle から RDF/XML への変換

下記 URL に Turtle 形式のデータを POST すると、RDF/XML 形式に変換されたデータが返される。

```
http://togows.org/convert/ttl.rdfxml
```

3.3.2. RDF/XML から Turtle への変換

下記 URL に RDF/XML 形式のデータを POST すると、Turtle 形式に変換されたデータが返される。

```
http://togows.org/convert/rdfxml.ttl
```

3.3.3. CSV 形式から RDF への変換

上記 URL に CSV 形式のデータを POST すると <format> で指定された形式に変換されたデータが返される。<format> には `rdxml` または `ttl` を指定できる。

```
http://togows.org/convert/csv.<format>
```

具体的には元データが

```
col1,col2
data11,data21
data12,data22
```

で `format` が `ttl` の場合、以下のようなデータが返される。

```
@prefix : <>
:entry1
  :col1 "data11" ;
  :col2 "data12" .
:entry2
  :col1 "data21" ;
  :col2 "data22" .
```

生成される Turtle 形式ではデータの可用性を考慮しあえて `prefix` が空白となっているため、このデータを `http://example.org/hoge/` で取れるようにウェブサイト上に置くと、`rapper`

コマンドを用いて下記のように（ウェブサイトの URL を `prefix` として）N-Triples 形式に変換することができる。

```
% rapper -i guess -o ntriples http://example.org/hoge/  
<http://example.org/hoge/entry1> <http://example.org/hoge/col1> "data11" .  
<http://example.org/hoge/entry1> <http://example.org/hoge/col2> "data12" .  
<http://example.org/hoge/entry2> <http://example.org/hoge/col1> "data21" .  
<http://example.org/hoge/entry2> <http://example.org/hoge/col2> "data22" .
```

3.3.4. GenBank 形式から RDF への変換

上記 URL に GenBank 形式のデータを POST すると `<format>` で指定された形式に変換されたデータが返される。 `<format>` には `ntriples`, `n3`, `rdxml`, `ttl` を指定できる。

```
http://togows.org/convert/genbank.<format>
```

3.3.5. EMBL 形式から RDF への変換

上記 URL に EMBL 形式のデータを POST すると `<format>` で指定された形式に変換されたデータが返される。 `<format>` には `ntriples`, `n3`, `rdxml`, `ttl` を指定できる。

```
http://togows.org/convert/embl.<format>
```

3.3.6. DDBJ 形式から RDF への変換

上記 URL に DDBJ 形式のデータを POST すると `<format>` で指定された形式に変換されたデータが返される。 `<format>` には `ntriples`, `n3`, `rdxml`, `ttl` を指定できる。

```
http://togows.org/convert/ddbj.<format>
```

3.3.7. HMMER3 table から RDF への変換

上記 URL に HMMER3 table 形式のデータを POST すると `<format>` で指定された形式に変換されたデータが返される。 `<format>` には `rdxml` を指定できる。

```
http://togows.org/convert/hmmer3tbl.<format>
```

4. TogoWS UCSC API

TogoWS では Ruby UCSC API の機能を利用して、UCSC で提供されている MySQL からのゲノムデータ取得を実現している。UCSC では様々な生物種（データベース）ごとに多数のテーブルが利用可能となっている。TogoWS では、これらの各データベースに対して下記の REST API で検索を行うことができる。

[http://togows.org/api/ucsc/データベース名/テーブル名/カラム名=値\[/offset,limit\]\[.json\]](http://togows.org/api/ucsc/データベース名/テーブル名/カラム名=値[/offset,limit][.json])

4.1. UCSC の対応データベース

データベースには、ヒトゲノム hg19、マウスゲノム mm10、酵母ゲノム sacCer2 などを指定できる。現状では、下記のデータベースがサポートされている。

- **human:** Hg38, Hg19, Hg18
- **mammals:** chimp (PanTro3), orangutan (PonAbe2), rhesus (RheMac2), marmoset (CalJac3), mouse (Mm10, Mm9), rat (Rn5, Rn4), guinea pig (CavPor3), rabbit (OryCun2), cat (FelCat4), panda (AilMel1), dog (CanFam2), horse (EquCab2), pig (SusScr2), sheep (OviAri1), cow (BosTau4), elephant (LoxAfr3), opossum (MonDom5), platypus (OrnAna1)
- **vertebrates:** chicken (GalGal3), zebra finch (TaeGut1), lizard (AnoCar2), X. tropicalis (XenTro2), zebrafish (DanRer7), tetraodon (TetNig2), fugu (Fr2), stickleback (GasAcu1), medaka (OryLat2), lamprey (PetMar1)
- **deuterostomes:** lancelet (BraFlo1), sea squirt (Ci2), sea urchin (StrPur2)
[insects] D.melanogaster (Dm3), D.simulans (DroSim1), D.sechellia (DroSec1), D.yakuba (DroYak2), D.erecta (DroEre1), D.ananassae (DroAna2), D.pseudoobscura (Dp3), D.persimilis (DroPer1), D.virilis (DroVir2), D.mojavensis (DroMoj2), D.grimshawi (DroGri1), Anopheles mosquito (AnoGam1), honey bee (ApiMel2)
- **nematodes:** C.elegans (Ce6), C.brenneri (CaePb3), C.briggsae (Cb3), C.remanei (CaeRem3), C.japonica (CaeJap1), P.pacificus (PriPac1)
- **others:** sea hare (AplCal1), yeast (SacCer2)
- **genome assembly independent:** Go, HgFixed, Proteome, UniProt, VisiGene

検索条件である「カラム名=値」はセミコロン ; で複数指定することができる。結果の取得件数は offset, limit で指定することができる。出力フォーマットには現状では省略時のテキスト形式のほか、JSON 形式が指定できる。

4.1.1. データベースの一覧

UCSC API の対応するデータベースの一覧は下記の REST API で取得することができる。

<http://togows.org/api/ucsc>

一覧を JSON 形式で取得するには下記の API を使用する。

<http://togows.org/api/ucsc.json>

```
["hg38", "hg19", "hg18", "panTro3", "ponAbe2", "rheMac2", "calJac3", "mm10", "mm9", "rn5", "rn4", "cavPor3", "oryCun2", "felCat4", "ailMel1", "canFam2", "equCab2", "susScr2", "oviAri1", "bosTau4", "loxAfr3", "monDom5", "ornAna1", "galGal3", "taeGut1", "anoCar2", "xenTro2", "danRer7", "tetNig2", "fr2", "gasAcu1", "oryLat2", "petMar1", "braFlo1", "ci2", "strPur2", "dm3", "droSim1", "droSec1", "droYak2", "droEre1", "droAna2", "dp3", "droPer1", "droVir2", "droMoj2", "droGri1", "anoGam1", "apiMel2", "ce6", "caePb3", "cb3", "caeRem3", "caeJap1", "priPac1", "aplCal1", "sacCer2", "go", "hgFixed", "proteome", "uniProt", "visiGene"]
```

4.1.2. テーブルの一覧

各データベースに含まれるテーブルはデータベースごとに異なるが、その一覧は下記の REST API で取得することができる。

[http://togows.org/api/ucsc/データベース名\[.json\]](http://togows.org/api/ucsc/データベース名[.json])

ヒトゲノム hg38 のテーブル一覧を JSON 形式で取得するには下記の API を使用する。

<http://togows.org/api/ucsc/hg38.json>

```
["all_est", "all_mrna", "altLocations", "altSeqLiftOverPsl", "author", "bigFiles", "cds", "cell", "chainCanFam3", "chainCanFam3Link", "chainMm10", "chainMm10Link", "chainRheMac3", "chainRheMac3Link", "chainSelf", "chainSelfLink", "chromInfo", "cpgIslandExt", "cpgIslandExtUnmasked", "ctgPos2", "cytoBandIdeo", "description", "development", "estOrientInfo", "extFile", "extNcbiAltSequence", "gap", "gbCdnaInfo", "gbExtFile", "gbLoaded", "gbMiscDiff", "gbSeq", "gbStatus", "gbWarn", "gc5BaseBw", "geneName", "genscan", "gold", "grp", "hg38ContigDiff", "hgFindSpec", "history", "imageClone", "intronEst", "keyword", "library", "mgcFullMrna", "mgcGenes", "mrna"]
```

```
Clone", "mrnaOrientInfo", "nestedRepeats", "netCanFam3", "netMm10", "netRheMac3",
"orfeomeGenes", "orfeomeMrna", "organism", "productName", "refFlat", "refGene", "r
efLink", "refSeqAli", "refSeqStatus", "refSeqSummary", "rmsk", "scaffolds", "seqNc
biAltSequence", "sex", "simpleRepeat", "source", "tableDescriptions", "tableList"
, "tissue", "trackDb", "ucscToINSDC", "wgEncodeGencode2wayConsPseudoV19", "wgEnco
deGencodeAnnotationRemarkV19", "wgEncodeGencodeAttrsV19", "wgEncodeGencodeBasi
cV19", "wgEncodeGencodeCompV19", "wgEncodeGencodeExonSupportV19", "wgEncodeGenc
odeGeneSourceV19", "wgEncodeGencodePdbV19", "wgEncodeGencodePolyaV19", "wgEncod
eGencodePseudoGeneV19", "wgEncodeGencodePubMedV19", "wgEncodeGencodeRefSeqV19"
, "wgEncodeGencodeTagV19", "wgEncodeGencodeTranscriptSourceV19", "wgEncodeGenco
deTranscriptSupportV19", "wgEncodeGencodeTranscriptionSupportLevelV19", "wgEnc
odeGencodeUniProtV19", "windowmaskerSdust", "xenoEst", "xenoMrna", "xenoRefFlat"
, "xenoRefGene", "xenoRefSeqAli"]
```

ヒトゲノムの場合、現状では hg19 から hg38 への移行フェーズであり、hg19 には 10993 ものテーブルが登録されている一方、hg38 ではまだ 98 テーブルしか提供されていないため、用途に応じて使い分ける必要がある。

4.1.3. カラムの一覧

データベース名、テーブル名を指定して検索を行う場合に、各テーブルのカラム名を利用するが、このカラム名の一覧は下記の REST API で取得することができる。

```
http://togows.org/api/ucsc/データベース名/テーブル名[.json]
```

具体的には、ヒトゲノム hg38 の refGene テーブルに含まれるカラム名の一覧は下記の API で取得できる。

```
http://togows.org/api/ucsc/hg38/refGene
```

結果として下記のようなデータが取得される。第1カラムがテーブルのカラム名で、第2カラムは各テーブルに含まれる値のサンプルである。

```
bin    646
name   NM_018416
chrom  chr12
strand      +
```

```

txStart      8032762
txEnd 8055522
cdsStart     8039832
cdsEnd       8052850
exonCount    11
exonStarts
    8032762,8039818,8042657,8043700,8043950,8044759,8047881,8048696,804936
1,8050521,8052761,
exonEnds
    8033833,8040165,8042732,8043769,8044091,8044958,8048289,8048798,804957
1,8050620,8055522,
score 0
name2 FOXJ2
cdsStartStat  cpl
cdsEndStat    cpl
exonFrames    -1,0,0,0,0,0,1,1,1,1,1,

```

同様の情報は下記の API により JSON 形式でも取得可能である。

<http://togows.org/api/ucsc/hg38/refGene.json>

```

[{"name":"bin","value":"646"}, {"name":"name","value":"NM_018416"}, {"name":"c
hrom","value":"chr12"}, {"name":"strand","value":"+"}, {"name":"txStart","valu
e":8032762}, {"name":"txEnd","value":8055522}, {"name":"cdsStart","value":8039
832}, {"name":"cdsEnd","value":8052850}, {"name":"exonCount","value":11}, {"nam
e":"exonStarts","value":"8032762,8039818,8042657,8043700,8043950,8044759,804
7881,8048696,8049361,8050521,8052761,"}, {"name":"exonEnds","value":"8033833,
8040165,8042732,8043769,8044091,8044958,8048289,8048798,8049571,8050620,8055
522,"}, {"name":"score","value":0}, {"name":"name2","value":"FOXJ2"}, {"name":"
cdsStartStat","value":"cpl"}, {"name":"cdsEndStat","value":"cpl"}, {"name":"
exonFrames","value":"-1,0,0,0,0,0,1,1,1,1,1,"}]

```

UCSC では数万を超える膨大な数のテーブルが提供されているため、各テーブルのスキーマについて把握することは困難であり、データベースやテーブル間でカラム名が統一されていないこともあるため、実際の値の例を見て検索対象のカラムを指定することが求められる。上記の第2カラムに載っている値のサンプルで分かりにくい場合には、各テーブル

から指定した件数のレコードを取得する下記の REST API を利用することができる。

[http://togows.org/api/ucsc/データベース名/テーブル名/offset,limit\[.json\]](http://togows.org/api/ucsc/データベース名/テーブル名/offset,limit[.json])

ヒトゲノム hg38 の refGene テーブルから最初の 10 件を取得するには下記の API を使用する。

<http://togows.org/api/ucsc/hg38/refGene/1,10>

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	name	chrom	strand	txStart	txEnd	cdsStart	cdsEnd	exonCount	exonStarts	exonEnds	score	name2	cdsStartStat	cdsEndStat	exonFrames
2	NM_018416	chr12	+	8032762	8055522	8039832	8052850	11	18,80489361,8050521,8052781,18,8049571,8050620,8055522,		0	FOXJ2	cmpl	cmpl	-1,0,0,0,0,1,1,1,1,1,
3	NR_024437	chr2_K1270767v1_alt	+	132762	158811	158811	158811	3	132762,138956,158708,	132871,139097,158811,	0	LOC728323	unk	unk	-1,-1,-1,
4	NR_038051	chr1	+	30365	30503	30503	30503	1	30365,	30503,	0	MIR1302-2	unk	unk	-1,
5	NM_005967	chr12	+	57088893	57095476	57089271	57094721	7	7093062,57093406,57094611,	7093195,57093988,57095476,	0	NAB2	cmpl	cmpl	0,2,0,2,0,1,1,
6	NM_002835	chr7	+	77537455	77640071	77537546	77639280	18	7637017,77638623,77639218,	7637048,77638731,77640071,	0	PTPN12	cmpl	cmpl	0,1,0,0,0,0,2,0,0,2,1,1,0,1,1,
7	NM_182697	chr7	-	129830732	129952960	129834936	129952555	5	79567,129880894,129952502,	79642,129880971,129952960,	0	UBE2H	cmpl	cmpl	1,1,1,2,0,
8	NM_004847	chr6_GL000252v2_alt	+	2862983	2864788	2863678	2864277	3	2862983,2863287,2863436,	2863060,2863349,2864788,	0	AIF1	cmpl	cmpl	-1,-1,0,
9	NR_024321	chr1	-	826205	827522	827522	827522	1	826205,	827522,	0	LINC00115	unk	unk	-1,
10	NM_021096	chr22	+	39570752	39689736	39570752	39686405	37	9682495,39684301,39685760,	9682661,39684498,39689736,	0	CACNA1I	cmpl	cmpl	0,2,0,0,0,1,0,2,1,0,1,0,0,0,0,1,0,
11	NM_021100	chr20	-	35668687	35699365	35669621	35699288	13	5697683,35698480,35699191,	5697800,35698590,35699365,	0	NFS1	cmpl	cmpl	2,2,1,0,1,1,0,0,0,0,1,0,

4.2. キーワード検索

UCSC ゲノムデータベースの検索は、上記からデータベース名、テーブル名、カラム名を指定して、カラムの値によるキーワード検索を下記の REST API によって行う。

[http://togows.org/api/ucsc/データベース名/テーブル名/カラム名=値\[/offset,limit\]\[.json\]](http://togows.org/api/ucsc/データベース名/テーブル名/カラム名=値[/offset,limit][.json])

[http://togows.org/api/ucsc/データベース名/テーブル名/カラム名!=値\[/offset,limit\]\[.json\]](http://togows.org/api/ucsc/データベース名/テーブル名/カラム名!=値[/offset,limit][.json])

具体例として、ヒトゲノム hg38 の refGene テーブルから name2 カラムに遺伝子名 UVSSA を含むものを検索する場合は下記の API を使用する。

<http://togows.org/api/ucsc/hg38/refGene/name2=UVSSA>

結果として下記のような表が返される。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	bin	name	chrom	strand	txStart	txEnd	cdsStart	cdsEnd	exonCount	exonStarts	exonEnds	score	name2	cdsStartStat	cdsEndStat	exonFrames
2	595	NM_020894	chr4	+	1347315	1388049	1348091	1385961	14	3879,1383765,1385967,3988,1383940,1388049,		0	UVSSA	cmpl	cmpl	1,1,0,0,1,2,2,0,1,2,

これを JSON 形式で取得するには下記の API を利用する。

<http://togows.org/api/ucsc/hg38/refGene/name2=UVSSA.json>

結果として下記の JSON データが取得できる。

```
[
  {
    "bin": "595",
    "name": "NM_020894",
    "chrom": "chr4",
    "strand": "+",
    "txStart": 1347315,
    "txEnd": 1388049,
    "cdsStart": 1348091,
    "cdsEnd": 1385961,
    "exonCount": 14,
    "exonStarts":
"1347315,1348089,1349523,1351714,1353029,1354734,1355116,1366319,1375363,1376
033,1380046,1380879,1383765,1385867,",
    "exonEnds":
"1347760,1348189,1349854,1351835,1353413,1354847,1355245,1366431,1375508,1376
168,1380230,1380988,1383940,1388049,",
    "score": 0,
    "name2": "UVSSA",
    "cdsStartStat": "cpl",
    "cdsEndStat": "cpl",
    "exonFrames": "-1,0,2,0,1,1,0,0,1,2,2,0,1,2,",
    "": null
  }
]
```

ヒトゲノムのバージョン **hg38** で **refGene** テーブルから染色体 4 番上に載っているアノテーションを 10 件取得するには下記の API を使用する。

<http://togows.org/api/ucsc/hg38/refGene/chrom=chr4/1,10>

結果として下記のような表が取得できる。

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	bin	name	chrom	strand	txStart	txEnd	cdsStart	cdsEnd	exonCount	exonStarts	exonEnds	score	name2	cdsStartStat	cdsEndStat	exonFrames
2	1	NM_001290	chr4	-	16501540	16898801	16502642	16898485	9	95702,16759157,16898353,95875,16759260,16898801,	0	LDB2	cmpl	cmpl	0,1,0,0,1,0,0,	
3	1	NM_018126	chr4	+	41935119	41960807	41935484	41954199	7	44792,41949301,41954069,44926,41949385,41960807,	0	TMEM33	cmpl	cmpl	0,0,2,1,0,2,2,	
4	1	NM_001130834	chr4	-	16501540	16898801	16505877	16898485	9	95702,16759157,16898353,95875,16759260,16898801,	0	LDB2	cmpl	cmpl	-1,0,1,0,0,1,0,0,	
5	1	NR_037934	chr4	+	25160671	25198505	25198505	25198505	2	25160671,25197467,	25160753,25198505,	0	SEPSECS-AS1	unk	unk	-1,-1,
6	1	NM_001101667	chr4	-	8366281	8440725	8370947	8416521	17	8415765,8416377,8440647,,8415999,8416535,8440725,	0	ACOX3	cmpl	cmpl	0,1,1,1,0,0,0,2,0,0,0,0,0,-1,	
7	1	NM_003501	chr4	-	8366281	8440725	8366960	8416521	18	8415765,8416377,8440647,,8415999,8416535,8440725,	0	ACOX3	cmpl	cmpl	0,1,1,1,0,0,0,2,0,0,0,0,0,-1,	
8	2	NM_032518	chr4	-	108823882	109302643	108824089	109302019	34	4582,109301722,109302168,1652,109302075,109302643,	0	COL25A1	cmpl	cmpl	1,0,0,0,0,0,0,0,0,0,0,1,0,0,-1,	
9	2	NM_001278538	chr4	-	75479036	75514430	75482537	75514264	9	108819,75509176,75514240,,08935,75509296,75514430,	0	RCHY1	cmpl	cmpl	0,2,2,0,0,2,0,0,0,	
10	2	NM_001278537	chr4	-	75479036	75514430	75482537	75514286	7	94100,755008819,75514196,,94179,755008935,75514430,	0	RCHY1	cmpl	cmpl	0,2,0,0,2,0,0,	
11	2	NM_001278536	chr4	-	75479036	75514430	75482537	75514286	8	94100,755008819,75514196,,94179,755008935,75514430,	0	RCHY1	cmpl	cmpl	0,2,2,0,0,2,0,0,	

複数のカラムに対する検索例としてヒトゲノムのバージョン hg19 で dbSNP のバージョン 138 から、染色体 22 番上でリファレンスの塩基が A (アデニン) であるアノテーションを 10 件、JSON 形式で取得する例は下記のようなになる。

<http://togows.org/api/ucsc/hg19/snp138/chrom=chr22;refUCSC=A/1,10.json>

結果として下記の SNP 情報が得られる。

```
[
  {
    "bin": "707",
    "chrom": "chr22",
    "chromStart": 16050035,
    "chromEnd": 16050036,
    "name": "rs374742143",
    "score": "0",
    "strand": "+",
    "refNCBI": "A",
    "refUCSC": "A",
    "observed": "A/C",
    "molType": "genomic",
    "class": "single",
    "valid": "by-cluster",
    "avHet": "0.0",
    "avHetSE": "0.0",
    "func": "unknown",
    "locType": "exact",
    "weight": 1,
    "exceptions": "",
    "submitterCount": "2",
```

```

"submitters": "BILGI_BIOE,GMI,",
"alleleFreqCount": "0",
"alleles": "",
"alleleNs": "",
"alleleFreqs": "",
"bitfields": "",
"": null
},

```

(省略)

4.3. 領域指定検索

実際には、キーワードや属性だけでなく、ゲノム配列上の領域を指定した検索も多く使われる。このため TogoWS では下記の REST API で領域の範囲を指定した検索が可能である。

<http://togows.org/api/ucsc/データベース名/テーブル名/配列名:領域-領域>

ここで、配列名には chr1 など生物種ごとに規定の染色体 ID を用い、領域には塩基位置の番号を指定する。塩基位置は 3 ケタごとにカンマ , で区切った形式も有効である。なお、TogoWS では生物学でよく利用される塩基番号を 1 から数える形式を利用しているが、UCSC の MySQL データベースでは塩基と塩基の間を 0 から数える形式を採用しているため、Ruby UCSC API の機能によって座標系を自動変換している。

具体的には、染色体 1 番の 100 万～110 万塩基対の領域にある遺伝子を refGene テーブルから取得するには下記の API を使用する。

<http://togows.org/api/ucsc/hg38/refGene/chr1:1,000,000-1,100,000>

I	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	bin	name	chrom	strand	txStart	txEnd	cdsStart	cdsEnd	exonCount	exonStarts	exonEnds	score	name2	cdsStartStat	cdsEndStat	exonFrames
2	9	NM_196576	chr1	+	1020122	1056119	1020172	1054981		36 1727,1053752,1054447,1054823,1815,1053977,1054551,1056119,		0	AGRN	cmpl	cmpl	.1,1,0,1,1,1,2,2,1,1,2,2,0,0,1,2,2,1,
3	592	NM_021170	chr1	-	998961	1000172	999058	999973	4	998961,999525,999691,999865,999432,999613,999787,1000172,		0	HES4	cmpl	cmpl	1,0,0,0,
4	592	NM_005101	chr1	+	1013466	1014539	1013573	1014478	2	1013466,1013983,	1013576,1014539,	0	ISG15	cmpl	cmpl	0,0,
5	592	NM_001142467	chr1	-	998963	1000172	999058	999973	3	998963,999525,999691,	999432,999613,1000172,	0	HES4	cmpl	cmpl	1,0,0,
6	593	NM_017891	chr1	-	1081817	1116356	1082892	1091543	10	3352,1091471,1091990,1116059,3428,1091565,1092103,1116356,		0	C1orf159	cmpl	cmpl	1,0,1,1,1,1,0,0,-1,-1,
7	593	NM_001205252	chr1	-	1071745	1074307	1071816	1072566	2	1071745,1074015,	1072575,1074307,	0	RNF223	cmpl	cmpl	0,-1,

同様に、染色体 1 番の 2 万～2 万 1 千塩基対の領域にある SNP をを JSON 形式で取得するには下記の API を使用する。

<http://togows.org/api/ucsc/hg19/snp138/chr1:20,000-21,000.json>

下記のような JSON が取得される。

```
[
  {
    "bin": "585",
    "chrom": "chr1",
    "chromStart": 20036,
    "chromEnd": 20037,
    "name": "rs12354133",
    "score": "0",
    "strand": "+",
    "refNCBI": "A",
    "refUCSC": "A",
    "observed": "A/C",
    "molType": "genomic",
    "class": "single",
    "valid": "unknown",
    "avHet": "0.0",
    "avHetSE": "0.0",
    "func": "unknown",
    "locType": "exact",
    "weight": 3,
    "exceptions": "MultipleAlignments",
    "submitterCount": "2",
    "submitters": "SC_SNP,SSAHASNP,",
    "alleleFreqCount": "0",
    "alleles": "",
    "alleleNs": "",
    "alleleFreqs": "",
    "bitfields": "",
    "": null
  },
  (省略)
```

4.3.1. 領域検索の境界指定

領域指定では、領域の境界にまたがるアノテーションが生じる。デフォルトでは境界領域

にまたがるアノテーションも含めて検索される。これを明示的に指定するには下記の REST API を使用する。

領域にまたがるアノテーションを含める場合 (デフォルト) :

<http://togows.org/api/ucsc/データベース名/テーブル名/配列名:領域-領域>

<http://togows.org/api/ucsc/データベース名/テーブル名/inclusive/配列名:領域-領域>

領域にまたがるアノテーションを除外する場合 :

<http://togows.org/api/ucsc/データベース名/テーブル名/exclusive/配列名:領域-領域>

染色体 4 番の 1,350,000~1,400,000 塩基対の領域には、1,347,315~1,388,049 にある UVSSA 遺伝子と 1,391,551~1,395,994 にある CRIPAK 遺伝子が含まれ、UVSSA 遺伝子の 5'端は境界にまたがって存在する。デフォルトの **inclusive** 検索では、このように境界にまたがる UVSSA 遺伝子も検索結果に含まれる。

<http://togows.org/api/ucsc/hg38/refGene/chr4:1,350,000-1,400,000.json>

<http://togows.org/api/ucsc/hg38/refGene/inclusive/chr4:1,350,000-1,400,000.json>

```
[{"bin": "595", "name": "NM_020894", "chrom": "chr4", "strand": "+", "txStart": 1347315, "txEnd": 1388049, "cdsStart": 1348091, "cdsEnd": 1385961, "exonCount": 14, "exonStarts": "1347315,1348089,1349523,1351714,1353029,1354734,1355116,1366319,1375363,1376033,1380046,1380879,1383765,1385867,", "exonEnds": "1347760,1348189,1349854,1351835,1353413,1354847,1355245,1366431,1375508,1376168,1380230,1380988,1383940,1388049,", "score": 0, "name2": "UVSSA", "cdsStartStat": "cpl", "cdsEndStat": "cpl", "exonFrames": "-1,0,2,0,1,1,0,0,1,2,2,0,1,2,", ""}, {"bin": "595", "name": "NM_175918", "chrom": "chr4", "strand": "+", "txStart": 1391551, "txEnd": 1395994, "cdsStart": 1394511, "cdsEnd": 1395852, "exonCount": 1, "exonStarts": "1391551,", "exonEnds": "1395994,", "score": 0, "name2": "CRIPAK", "cdsStartStat": "cpl", "cdsEndStat": "cpl", "exonFrames": "0,", ""}]
```

これを **exclusive** 検索に変更すると境界にまたがる UVSSA 遺伝子のアノテーションが除外される。

<http://togows.org/api/ucsc/hg38/refGene/exclusive/chr4:1,350,000-1,400,000.json>

```
[{"bin": "595", "name": "NM_175918", "chrom": "chr4", "strand": "+", "txStart": 1391551, "txEnd": 1395994, "cdsStart": 1394511, "cdsEnd": 1395852, "exonCount": 1, "exonStarts": "1391551", "exonEnds": "1395994", "score": 0, "name2": "CRIPAK", "cdsStartStat": "cpl", "cdsEndStat": "cpl", "exonFrames": "0", "": null}]
```

4.4. CDS, exon, intron の検索

各遺伝子の CDS 領域、exon 領域、intron 領域のリストを得るため、refGene など遺伝子アノテーションのテーブルでは下記の REST API が利用できる。

<http://togows.org/api/ucsc/hg38/refGene/カラム名=値/cdss.json>

<http://togows.org/api/ucsc/hg38/refGene/カラム名=値/exons.json>

<http://togows.org/api/ucsc/hg38/refGene/カラム名=値/introns.json>

具体的には UVSSA 遺伝子の場合、CDS 領域の位置のリストは、

<http://togows.org/api/ucsc/hg38/refGene/name2=UVSSA/cdss.json>

```
[{"chrom": "chr4", "chr_start": 1348092, "chr_end": 1348189}, {"chrom": "chr4", "chr_start": 1349524, "chr_end": 1349854}, {"chrom": "chr4", "chr_start": 1351715, "chr_end": 1351835}, {"chrom": "chr4", "chr_start": 1353030, "chr_end": 1353413}, {"chrom": "chr4", "chr_start": 1354735, "chr_end": 1354847}, {"chrom": "chr4", "chr_start": 1355117, "chr_end": 1355245}, {"chrom": "chr4", "chr_start": 1366320, "chr_end": 1366431}, {"chrom": "chr4", "chr_start": 1375364, "chr_end": 1375508}, {"chrom": "chr4", "chr_start": 1376034, "chr_end": 1376168}, {"chrom": "chr4", "chr_start": 1380047, "chr_end": 1380230}, {"chrom": "chr4", "chr_start": 1380880, "chr_end": 1380988}, {"chrom": "chr4", "chr_start": 1383766, "chr_end": 1383940}, {"chrom": "chr4", "chr_start": 1385868, "chr_end": 1385961}]
```

exon 領域の位置のリストは、

<http://togows.org/api/ucsc/hg38/refGene/name2=UVSSA/exons.json>

```
[{"chrom": "chr4", "chr_start": 1347316, "chr_end": 1347760}, {"chrom": "chr4", "chr_start": 1348090, "chr_end": 1348189}, {"chrom": "chr4", "chr_start": 1349524, "chr_end": 1349854}, {"chrom": "chr4", "chr_start": 1351715, "chr_end": 1351835}, {"chrom": "chr4", "chr_start": 1353030, "chr_end": 1353413}, {"chrom": "chr4", "chr_start": 1354735, "chr_end": 1354847}, {"chrom": "chr4", "chr_start": 1355117, "chr_end": 1355245}, {"chrom": "chr4", "chr_start": 1366320, "chr_end": 1366431}, {"chrom": "chr4", "chr_start": 1375364, "chr_end": 1375508}, {"chrom": "chr4", "chr_start": 1376034, "chr_end": 1376168}, {"chrom": "chr4", "chr_start": 1380047, "chr_end": 1380230}, {"chrom": "chr4", "chr_start": 1380880, "chr_end": 1380988}, {"chrom": "chr4", "chr_start": 1383766, "chr_end": 1383940}, {"chrom": "chr4", "chr_start": 1385868, "chr_end": 1385961}]
```

```
, {"chr_start":1375364,"chr_end":1375508}, {"chrom":"chr4","chr_start":1376034,"chr_end":1376168}, {"chrom":"chr4","chr_start":1380047,"chr_end":1380230}, {"chrom":"chr4","chr_start":1380880,"chr_end":1380988}, {"chrom":"chr4","chr_start":1383766,"chr_end":1383940}, {"chrom":"chr4","chr_start":1385868,"chr_end":1388049}]
```

intron 領域の位置のリストは、

<http://togows.org/api/ucsc/hg38/refGene/name2=UVSSA/introns.json>

```
[{"chrom":"chr4","chr_start":1347761,"chr_end":1348089}, {"chrom":"chr4","chr_start":1348190,"chr_end":1349523}, {"chrom":"chr4","chr_start":1349855,"chr_end":1351714}, {"chrom":"chr4","chr_start":1351836,"chr_end":1353029}, {"chrom":"chr4","chr_start":1353414,"chr_end":1354734}, {"chrom":"chr4","chr_start":1354848,"chr_end":1355116}, {"chrom":"chr4","chr_start":1355246,"chr_end":1366319}, {"chrom":"chr4","chr_start":1366432,"chr_end":1375363}, {"chrom":"chr4","chr_start":1375509,"chr_end":1376033}, {"chrom":"chr4","chr_start":1376169,"chr_end":1380046}, {"chrom":"chr4","chr_start":1380231,"chr_end":1380879}, {"chrom":"chr4","chr_start":1380989,"chr_end":1383765}, {"chrom":"chr4","chr_start":1383941,"chr_end":1385867}]
```

のように得ることができる。

なお、JSON 形式を指定しなかった場合はテキスト形式で表示される。

<http://togows.org/api/ucsc/hg38/refGene/name2=UVSSA/exons>

```
chr4:1347316-1347760 chr4:1348090-1348189 chr4:1349524-1349854
chr4:1351715-1351835 chr4:1353030-1353413 chr4:1354735-1354847
chr4:1355117-1355245 chr4:1366320-1366431 chr4:1375364-1375508
chr4:1376034-1376168 chr4:1380047-1380230 chr4:1380880-1380988
chr4:1383766-1383940 chr4:1385868-1388049
```

4.5. bigWig ファイルの検索

UCSC からは bigWig データと関連するテーブルも提供されている。TogoWS では、これらのテーブルにアクセスされたタイミングで bigWig ファイルの自動取得を行いキャッシュする。また、bigWigInfo コマンドの結果、bigWigSummary コマンドの結果を得るための REST API を提供している。bigWig テーブルに関する情報は下記の API で取得できる。

<http://togows.org/api/ucsc/データベース名/bigWigテーブル名>

例として、hg19 の wgEncodeBroadHistoneGm12878H3k27acStdSig テーブルの場合、

<http://togows.org/api/ucsc/hg19/wgEncodeBroadHistoneGm12878H3k27acStdSig>

を指定すると、

```
fileName
    /gbdb/hg19/bbi/wgEncodeBroadHistoneGm12878H3k27acStdSig.bigWig
```

のように 1 カラムのテーブルにファイル名が格納されていることが確認できる。

bigWig ファイルに対する bigWigInfo コマンドの出力は、下記の REST API で取得できる。

<http://togows.org/api/ucsc/データベース名/bigWigテーブル名/info>

上記のテーブルの場合、

```
http://togows.org/api/ucsc/hg19/wgEncodeBroadHistoneGm12878H3k27acStdSig/info
version: 4
isCompressed: yes
isSwapped: 0
primaryDataSize: 198,894,024
primaryIndexSize: 1,440,088
zoomLevels: 10
chromCount: 23
basesCovered: 1,145,311,185
mean: 3.163029
min: 0.040000
max: 223899.000000
std: 98.594295
```

のような情報が表示される。

さらに、指定した領域の平均値(デフォルト)、最小値、最大値、標準偏差、カバレッジを下記の REST API で取得することが可能である。値を取得する領域は「配列名:領域 1-領域 2」の形式で指定し、この領域を何分割して各分割領域の値を取得するかを分割数で指定する必要がある。

<http://togows.org/api/ucsc/データベース名/bigWig テーブル名/領域/分割数>

bigWigSummary では、デフォルトで平均値が表示される。

<http://togows.org/api/ucsc/hg19/wgEncodeBroadHistoneGm12878H3k27acStdSig/chr1:100000-200000/1>
1.05991

ここで分割数を 5 に変更すると、10 万塩基の領域を 2 万塩基ごとに分割して得られた平均値をそれぞれ取得できる。

<http://togows.org/api/ucsc/hg19/wgEncodeBroadHistoneGm12878H3k27acStdSig/chr1:100000-200000/5>
1.07461 1.12662 1.03598 0.9346 n/a

値の計算できなかった領域については n/a が表示される。

平均値を明示的に指定する場合には mean を、最小値を指定するには min を、最大値を指定するには max を、標準偏差を指定するには std を、カバレッジを指定するには coverage をそれぞれ下記の REST API で指示する。

<http://togows.org/api/ucsc/データベース名/bigWig テーブル名/領域/分割数/mean>
<http://togows.org/api/ucsc/データベース名/bigWig テーブル名/領域/分割数/min>
<http://togows.org/api/ucsc/データベース名/bigWig テーブル名/領域/分割数/max>
<http://togows.org/api/ucsc/データベース名/bigWig テーブル名/領域/分割数/std>
<http://togows.org/api/ucsc/データベース名/bigWig テーブル名/領域/分割数/coverage>

平均値

<http://togows.org/api/ucsc/hg19/wgEncodeBroadHistoneGm12878H3k27acStdSig/chr1:100000-200000/5/mean>
1.07461 1.12662 1.03598 0.9346 n/a

最小値

```
http://togows.org/api/ucsc/hg19/wgEncodeBroadHistoneGm12878H3k27acStdSig/chr
1:100000-200000/5/min
0.08 0.04 0.04 0.04 n/a
```

最大値

```
http://togows.org/api/ucsc/hg19/wgEncodeBroadHistoneGm12878H3k27acStdSig/chr
1:100000-200000/1/max
3 2 2 2 n/a
```

標準偏差

```
http://togows.org/api/ucsc/hg19/wgEncodeBroadHistoneGm12878H3k27acStdSig/chr
1:100000-200000/5/std
0.564019 0.502186 0.391216 0.347517 n/a
```

カバレッジ

```
http://togows.org/api/ucsc/hg19/wgEncodeBroadHistoneGm12878H3k27acStdSig/chr
1:100000-200000/5/coverage
0.15225 0.06935 0.062 0.1248 n/a
```

4.6. bigBed ファイルの検索

UCSC からは **bigWig** データのほかに、**bigBed** と関連するテーブルも提供されている。TogoWS では、これらのテーブルにアクセスされたタイミングで **bigBed** ファイルの自動取得を行いキャッシュする。また、**bigBedInfo** コマンドの結果、**bigBedSummary** コマンドの結果を得るための REST API を提供している。**bigBed** テーブルに関する情報は下記の API で取得できる。

```
http://togows.org/api/ucsc/データベース名/bigBed テーブル名
```

例として、hg19 の **wgEncodeDukeAffyExonA549SimpleSignalRep1** テーブルの場合、

```
http://togows.org/api/ucsc/hg19/wgEncodeDukeAffyExonA549SimpleSignalRep1
```

を指定すると、

```
fileName
    /gdbdb/hg19/bbi/wgEncodeDukeAffyExonA549SimpleSignalRep1.bigBed
```

のように 1 カラムのテーブルにファイル名が格納されていることが確認できる。

bigBed ファイルに対する **bigBedInfo** コマンドの出力は、下記の REST API で取得できる。

<http://togows.org/api/ucsc/データベース名/bigBed テーブル名/info>

上記のテーブルの場合、

```
http://togows.org/api/ucsc/hg19/wgEncodeDukeAffyExonA549SimpleSignalRep1/info
version: 4
hasHeaderExtension: no
isCompressed: yes
isSwapped: 0
extraIndexCount: 0
itemCount: 38,378
primaryDataSize: 796,830
primaryIndexSize: 6,900
zoomLevels: 7
chromCount: 24
basesCovered: 1,143,378,960
meanDepth (of bases covered): 1.055693
minDepth: 1.000000
maxDepth: 18.000000
std of depth: 0.310857
```

のような情報が表示される。

さらに、指定した領域の平均値、最小値、最大値、カバレッジ(デフォルト)を下記の REST API で取得することが可能である。値を取得する領域は「配列名:領域 1-領域 2」の形式で指定し、この領域を何分割して各分割領域の値を取得するかを分割数で指定する必要がある。

<http://togows.org/api/ucsc/データベース名/bigBed テーブル名/領域/分割数>

bigBedSummary では、デフォルトでカバレッジが表示される。

```
http://togows.org/api/ucsc/hg19/wgEncodeDukeAffyExonA549SimpleSignalRep1/chr1:
2000000-3000000/1
0.505341
```

ここで分割数を 5 に変更すると、100 万塩基の領域を 20 万塩基ごとに分割して得られた平均値をそれぞれ取得できる。

```
http://togows.org/api/ucsc/hg19/wgEncodeDukeAffyExonA549SimpleSignalRep1/chr1:
2000000-3000000/5
0.926075 0.62637 0.41576 0.53711 0.04999
```

値の計算できなかった領域については n/a が表示される。

平均値を指定する場合には mean を、最小値を指定するには min を、最大値を指定するには max を、カバレッジ明示的に指定するには coverage をそれぞれ下記の REST API で指示する。

```
http://togows.org/api/ucsc/データベース名/bigBed テーブル名/領域/分割数/mean
http://togows.org/api/ucsc/データベース名/bigBed テーブル名/領域/分割数/min
http://togows.org/api/ucsc/データベース名/bigBed テーブル名/領域/分割数/max
http://togows.org/api/ucsc/データベース名/bigBed テーブル名/領域/分割数/coverage
```

平均値

```
http://togows.org/api/ucsc/hg19/wgEncodeDukeAffyExonA549SimpleSignalRep1/chr1:
2000000-3000000/5/mean
1.02093 1.01923 1 1 1.0209
```

最小値

```
http://togows.org/api/ucsc/hg19/wgEncodeDukeAffyExonA549SimpleSignalRep1/chr1:
2000000-3000000/5/min
1 1 1 1 1
```

最大値

```
http://togows.org/api/ucsc/hg19/wgEncodeDukeAffyExonA549SimpleSignalRep1/chr1:
2000000-3000000/5/max
```

```
3 2 1 1 2
```

カバレッジ

```
http://togows.org/api/ucsc/hg19/wgEncodeDukeAffyExonA549SimpleSignalRep1/chr1:
2000000-3000000/5/coverage
```

```
0.926075 0.62637 0.41576 0.53711 0.04999
```

4.7. 2bit ファイルからの塩基配列の取得

UCSC ではゲノム配列は 2bit ファイルとして別途配布されている。TogoWS ではこれを取
得し、Ruby UCSC API を用いて塩基配列の取得を行う下記の REST API を提供している。

```
http://togows.org/api/ucsc/データベース名/配列名:領域 1-領域 2[.fasta]
```

具体的にヒトゲノム hg38 の染色体 1 番から 12345~12456 塩基の配列を取得するには下記
の API を使用する。

```
http://togows.org/api/ucsc/hg38/chr1:12,345-12,456
```

```
TCAGACCAGCCGGCTGGAGGGAGGGGCTCAGCAGGTCTGGCTTTGGCCCTGGGAGAGCAGGTGGAAGATCAGGCAGG
CCATCGCTGCCACAGAACCCAGTGGATTGGCCTAG
```

領域 1 の数値が領域 2 の数値よりも大きい場合、相補鎖の塩基配列が返される。

```
http://togows.org/api/ucsc/hg38/chr1:12,456-12,345
```

```
ctaggccaatccactgggttctgtggcagcgatggcctgctgatcttccacctgctctcccagggccaaagccag
acctgctgagcccctccctccagccggctggtctga
```

また、拡張子 .fasta を指定することにより FASTA 形式で配列を取得することが可能であ
る。

```
http://togows.org/api/ucsc/hg38/chr1:12345-12456.fasta
```

```
>hg38:chr1:12345-12456
```

```
TCAGACCAGCCGGCTGGAGGGAGGGGCTCAGCAGGTCTGGCTTTGGCCCTGGGAGAGCAG
GTGGAAGATCAGGCAGGCCATCGCTGCCACAGAACCCAGTGGATTGGCCTAG
```