

Sample SQL queries for cisRED Human_8, Mouse_2 and Rat_1 databases

- + MySQL database access: URL: db.cisred.org. Username: anonymous. Password should be left blank.
- + Human, mouse and rat schemas are identical.
- + Schemas are described by a 'Schema diagram' and a 'Database Tables and Fields' document, which are available from the 'Databases & Methods' page.
- + The example queries below are shown for the Human_8 database.

You can find answers to different kinds of biological questions by querying cisRED databases. Here we show SQL queries for some common questions. The examples may be directly useful to you, and should also help you construct other queries. If you have questions, please email us at cisred@bcgsc.ca.

Finding motifs and its features for a specific gene

You can easily query the 'features' table for information about conserved motifs discovered for the promoter region for a specific gene. The 'features' table holds results only for the database's 'target' species (here, human); the next query shows how to get information about a motif's orthologous species. Here we query Ensembl gene id ENSG00000001036.

```
SELECT
id AS motif, seqname AS chromosome, start, end, strand AS st, consensus
FROM
features
WHERE
ensembl_gene_id='ENSG00000001036';
```

The query results show that this gene has five conserved motifs: craHsap6, craHsap66, ...

motif	chromosome	start	end	st	consensus
6	6	143876927	143876936	-	ATATTTTnT
66	6	143876850	143876857	-	nmATATTT
23	6	143876046	143876051	-	ACAAAA
410	6	143875503	143875518	-	TTCTTAAACrYyCATT
452	6	143874551	143874569	-	TTGGTCAGGTGACmGCCGC

Finding motifs and its features for a specific gene along with its orthogous species

The previous query displays conserved motif features only for the target species (human). To also display sequences for orthologous species in a conserved motif, we need to include both 'features' and 'siterequences' tables. Here again we display motifs for gene 'ENSG00000001036'.

```
SELECT
s.feature_id AS motif, s.source_chromosome AS chromosome, s.source_start AS start,
s.source_end AS end, s.source_strand AS st, s.source_annotation AS gene, s.sequence,
s.tax
FROM
features f, sitesequences s
WHERE
f.id=s.feature_id AND f.ensembl_gene_id='ENSG00000001036'
ORDER BY
s.feature_id;
```

In the query results, you can distinguish sequence sets for each of the five motifs by the IDs in the first column. For example, motif craHsap6 is conserved across four species, and motif craHsap452 across seven.

motif	chromosome	start	end	st	gene	sequence	tax
6	5	146792269	146792278	-	ENSG00000001036_PTRO	ATATTTTGT	9598
6	SCAFFOLD60027	108190	108199	-	ENSG00000001036_MMUL	ATATTTTCCC	9544
6	scaffold_13303	19725603	19725612	-	ENSG00000001036_MDOM	ATATTTTTTT	13616
6	6	143876927	143876936	-	ENSG00000001036	ATATTTTGT	9606
66	SCAFFOLD60027	108194	108201	-	ENSG00000001036_MMUL	TCATATTT	9544
66	scaffold_13303	19725607	19725614	-	ENSG00000001036_MDOM	CAATATTT	13616
66	6	143876850	143876857	-	ENSG00000001036	AAATATTT	9606

66	5	146792273	146792280	-	ENSG00000001036_PTRO	CCATATTT	9598
223	1	8289312	8289317	-	ENSG00000001036_RNOR	ACAAAA	10116
223	SCAFFOLD60027	109445	109450	-	ENSG00000001036_MMUL	ACAAAA	9544
223	scaffold_13303	19725601	19725606	+	ENSG00000001036_MDOM	ACAAAA	13616
223	ENSG00000001036_MLUC	547	552	-	ENSG00000001036_MLUC	ACAAAA	59463
223	3	51375691	51375696	-	ENSG00000001036_GGAL	ACAAAA	9031
223	1	38043527	38043532	-	ENSG00000001036_CFAM	ACAAAA	9615
223	6	143876046	143876051	-	ENSG00000001036	ACAAAA	9606
223	10	13210743	13210748	+	ENSG00000001036_MMUS	ACAAAA	10090
223	5	146792269	146792274	+	ENSG00000001036_PTRO	ACAAAA	9598
410	5	146790852	146790867	-	ENSG00000001036_PTRO	TTCTTAAACCACTGGT	9598
410	6	143875503	143875518	-	ENSG00000001036	TTCTTAAACATTCATT	9606
410	SCAFFOLD60027	108468	108483	-	ENSG00000001036_MMUL	TTCTTAAACATTCATT	9544
410	10	13211304	13211319	+	ENSG00000001036_MMUS	TTCTTAAACATTCATT	10090
410	1	8288913	8288928	+	ENSG00000001036_RNOR	TTCTTAAATGCCCATTCATT	10116
410	scaffold_13303	19728539	19728554	+	ENSG00000001036_MDOM	TTCTTAAAAGCATCTT	13616
452	10	13211902	13211920	-	ENSG00000001036_MMUS	TTGGTCAGATGACTACTGT	10090
452	1	8289533	8289551	+	ENSG00000001036_RNOR	TTGGTCAGGTGACAGGAGC	10116
452	5	146789913	146789931	-	ENSG00000001036_PTRO	TTGGTCAGGTGACCGCGGC	9598
452	scaffold_13303	19728785	19728803	+	ENSG00000001036_MDOM	TGAGTCAGGTGACAGACTA	13616
452	SCAFFOLD60027	108004	108022	-	ENSG00000001036_MMUL	TTGGTCAGGTGACAGCGGC	9544
452	6	143874551	143874569	-	ENSG00000001036	TTGGTCAGGTGACCGCGGC	9606
452	1	38042700	38042718	-	ENSG00000001036_CFAM	GCCGTCAGGTGACCGCGGC	9615

Finding IDs for a motif that belongs to a specific group of similar motifs

In cisRED, motifs are grouped 'de novo' by using a pairwise edit distance metric, followed by OPTICS hierarchical clustering and BOSS segmentation (see [NAR 2006](#)). Here we search for motifs belonging to de novo group crHsap19358.

```
SELECT
feature_id AS motif
FROM
group_content
WHERE
group_id=19358;
```

The de novo group contains four conserved motifs:

```
motif
2955
403120
287423
328247
```

Finding known transcription factor binding site models that are similar to a specific motif

We annotate cisRED motifs using known transcription factor models in, e.g. the JASPAR database. The results are stored in the feature_similarity table. Here, we search for all JASPAR transcription factor binding site models that have an annotation p-value less than 0.0025 for motif craHsap29316:

```
SELECT
name
FROM
feature_similarity
WHERE
feature_id=29316
AND
pvalue < 0.0025;
```

Given the annotation p-value filter, the query returns only one JASPAR model:

name
TCF1

Finding all co-occurring patterns for a motif

After we identify *de novo* groups of similar motifs, we identify sets of group labels that co-occur within genomic regions of e.g. 200 nt in many search regions. We call such a co-occurring motif set a 'pattern', and call each co-occurrence of a pattern a 'pattern instance'. Here we ask whether motif *craHsap29316* is found in any *de novo* co-occurring motif patterns.

```
SELECT
pi.pattern_id, pi.region_id
FROM
pattern_instance_content pic, pattern_instance pi
WHERE
pic.pattern_instance_id=pi.id
AND
pic.feature_id=29316;
```

Motif *craHsap29316* occurs in four patterns with one search region:

pattern_id	region_id
16	7395
369	7395
2251	7395
2282	7395

Finding motifs and their genes for a co-occurring pattern

If we want to know not only the motifs in a pattern, but also which gene promoter regions have this pattern, we need to query the *search_region* table as well as the co-occurrence tables. Here we search for all features and genes for pattern *crmHsap16*.

```
SELECT
DISTINCT(pic.feature_id) AS motif, s.ensembl_gene_id AS gene
FROM
search_region s, pattern_instance pi, pattern_instance_content pic
WHERE
pi.pattern_id=16
AND
pi.id=pic.pattern_instance_id
AND
pi.region_id=s.id;
```

Pattern *crmHsap16* is present in two gene promoter regions:

motif	gene
328282	ENSG00000172264
328516	ENSG00000172264
28692	ENSG00000162692
29091	ENSG00000162692
29316	ENSG00000162692

Find details for all motifs in a pattern instance in a specific search region

To extend the previous query to give more detailed information about a motif's features (e.g. coordinates, consensus sequence, etc.), we add the *features* table to the query. Here we search for information on all motifs in pattern *crmHsap16*'s pattern_instances in search region 7395. By not filtering on motif discovery p-value, we accept its default value.

```
SELECT
f.id AS motif, f.seqname AS chromosome, f.start, f.end, f.strand AS st, f.consensus
FROM pattern_instance pi, pattern_instance_content pic, features f, search_region s
WHERE
f.id=pic.feature_id
```

```

AND
pi.id=pic.pattern_instance_id
AND
s.id=pi.region_id
AND
pi.region_id=7395
AND
pi.pattern_id=16;

```

motif	chromosome	start	end	st	consensus
28692	1	100896773	100896784	+	TTTATGAATAAA
29091	1	100896833	100896840	+	AAAAGAAA
29316	1	100896869	100896879	+	rAAATwATTy

Finding all annotation-based patterns for a motif

In addition to cluster-based (*de novo*) patterns, we also identify annotation-based patterns. These are co-occurring patterns of motif group labels, where each group label is the known transcription factor binding site model that gives the lowest annotation p-value. For example, when motifs that annotate best against models for NFIL3 and Pbx may co-occur in the pair pattern NFIL3__Pbx. Here, we display patterns and genes for motif 358498. By not filtering on either discovery p-value or annotation p-value, we accept their default values.

```

SELECT
pattern_name, ensembl_gene_id AS gene
FROM
annotation_patterns
WHERE
feature_id=358498;

```

Motif **craHsap358498** is present in four distinct co-occurring pair patterns in the promoter search region of one gene:

pattern_name	gene
NFIL3__Pbx	ENSG00000198756
En1__Pbx	ENSG00000198756
En1__NFIL3	ENSG00000198756
Pbx__Pbx	ENSG00000198756