Scheduled Mass Submission
(based on JPred REST API)
Intro

• Scheduled Mass-Submission is based on JPred RESTful API
  – Examples of simple scripts are given on the page
• Below we present details on a more advanced script, which is extremely user friendly and with a push of a button it allows to...
  – Run virtually unlimited number of sequences through JPred4/JNet protein secondary structure prediction (mass submission)
  – In a controlled way (scheduling)
  – Retrieving results of interest to a local directory (fully automatic submission/status check/retrieval)
Input is...

- Either a FASTA format file with multiple entries you’d like to submit (see example below)
- Or individual sequence FASTA files

Example of multiple entries FASTA file

```
>IP_057216.1.OR4F5_1
MQAPTLHYNYVWQRMCRHYGCHMGNWLSPFGEPVGVCRAŁTLWŚQ
>IP_057218.1.OR4F29 OR4F16 OR4F3 OR4F21_1,5,8
MTCSdaleASASHAASKSSSTSLLVWWRCSCS
>IP_057219.1.OR4F29 OR4F16 OR4F3_1,5
MGDPAPPSPVFLCAŁCİGHKYWKHPHCWFCĐH
>IP_057220.1.SAMD11_1
MEGATPLPVKLPŠRRMGPLWFQGPPTPISLČEVAGGRGGATQISRHSTQWČPLPPILYFRFRMQLKFK
>IP_057221.1.SAMD11_1
MARKTSPPKTRTERPPRQLLGGAGPILRAKLQLEGPAKRGRGFSQGPČHWASŁMPSAPTSTQR
>IP_057222.1.SAMD11_1
MGRRPQPLRTSPSGPWMTSAAŚVAGŁĂVESTŁGœurSGSTGRPČHC
```
Scheduling intro

- **Controlled submission** (scheduling) is important when running **A LOT of sequences**
  - JPred server is connected to a cluster where it is allowed to run in parallel up to 50-70 jobs
  - Which means if one user submits e.g. 1000 jobs in one go - any other user’s submission would get into queuing and would have to wait for a while before automatically submitted (same is true for the majority of the 1000 jobs submission - most of those jobs will be pending for a while before submission)
  - **HOWEVER**: submitting in a controlled way, not more than \( N \) jobs at a time but **continuously** (we recommend \( N \) around 20-40) - would allow any other user to still submit a job and get results quickly
  - **EXAMPLE on the next page will make it clear**
Submission Examples

**LEFT figure:** one can see a large (batch) submission: immediately a lot of job queuing, as JPred is limited to maximum 50-70 jobs running in parallel (depending on the cluster load).

**RIGHT figure:** example of controlled mass submission using scheduling (through JPred RESTful API). In this example a user controls submissions (automatically) to not more than 40 jobs, i.e. below the max number of nodes available to JPred jobs. Therefore any other user who would like to submit a job - would be able to do it without longtime waiting.
Step-by-step

Next slides go through all the steps of running the scheduled mass-submission.
Download code

- **Click download URL** or use e.g. ‘wget’ as below

Resolving www.compbio.dundee.ac.uk... 10.0.3.23
Connecting to www.compbio.dundee.ac.uk|10.0.3.23|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 7709 (7.5K) [application/x-gzip]
Saving to: `jpredMassSubmitSchedule.tar.gz'

100%[==============================================>] 7,709  --.-K/s  in  0s

2015-05-14 13:52:42 (268 MB/s) - `jpredMassSubmitSchedule.tar.gz' saved [7709/7709]

14/05/15 13:52:42 ~/test >
14/05/15 13:52:44 ~/test > ls -ltrh
total 0
-rw-r--- 1 adrozdetskiy lsd 7.6K May 14 13:47 jpredMassSubmitSchedule.tar.gz
14/05/15 13:52:48 ~/test >
14/05/15 13:52:49 ~/test >
14/05/15 13:52:49 ~/test > tar xzf jpredMassSubmitSchedule.tar.gz
14/05/15 13:52:54 ~/test >
14/05/15 13:52:54 ~/test >
14/05/15 13:52:55 ~/test > ls -ltrh
total 16K
-rwxr-xr-x 1 adrozdetskiy lsd 15K Mar 24 10:16 jpredapi
-rw-r--- 1 adrozdetskiy lsd 472 May 14 10:18 test_seqs.fa
-rwx------ 1 adrozdetskiy lsd 6.6K May 14 13:20 massSubmitScheduler.csh
-rwx------ 1 adrozdetskiy lsd 2.7K May 14 13:39 prepareInputs.csh
-rwx------ 1 adrozdetskiy lsd 7.6K May 14 13:47 jpredMassSubmitSchedule.tar.gz
14/05/15 13:52:58 ~/test >

Check you got the file

Open archive with ‘tar xzf’ or your favorite archive program.

Check you got the scripts (jpredapi, massSubmitScheduler.csh, and prepareInputs.csh) and test input file.
Prepare inputs

Have the sequences you would like to submit to JPred in FASTA file format, as in the example below.

```
14/05/15 13:53:00 ~/test > cat test_seqs.fa
>IP_057216.1_OR4F5_1
MQAPTLHYNYVWQRMCRHYGCHMGNWLSPFEGEPVGVCRAALTLWSQ
>IP_057218.1_OR4F29 OR4F16 OR4F3 OR4F21_1,5,8
MTCSAKSSPLEAASLKSSSTSSTSLVVWRWCCS
>IP_057219.1_OR4F29 OR4F16 OR4F3_1,5
MGDPPAPPSVFCLACGKHYWKHOPHCEVFDH
>IP_057220.1_SAMD11_1
MEGATPLPVKLPSSRMGPWLYFQGPPPTLPSLCVEVAGGRGGATQISRSHHSTQWPCLPPLYFFRFRMQNKKF
>IP_057221.1_SAMD11_1
MARKTPPKTRTERTPRQQLGAGGPLRALKLQLEGPAPRGRGFSQGPHCPWALSMPASPSTQAR
>IP_057222.1_SAMD11_1
MGRRPQPLRTSPSGPWMTSAASWGACLAVESTLGSNGSRGSTGPRPCCH
```

Note that first word of every sequence annotation should be a unique identifier, UniProt ID e.g.

Alternatively, have a set of individual sequence FASTA files (i.e. skipping the STEP 1 - next page)
STEP 1: run `prepareInputs.csh`

JPred REST API: prepare input FASTA files script (v.1)
Author: Dr. Alexey Drozdetskiy
Email: a.drozdetskiy@dunde.ac.uk
The Barton Group, Division of Computational Biology
College of Life Sciences, University of Dundee
www.compbio.dundee.ac.uk

Please contact me with suggestions to improve/fix client/server part of the JPred API and the mass submission/scheduling scripts.

Usage:

```bash
./prepareInputs.csh filename
```

where filename - is the name of the local file with multiple FASTA format entries you would want to submit as individual sequences.

**NOTE:** multiple controlled individual sequence submission through the JPred REST API is a preferred way over single batch submission. For details – please see tutorials on mass submission (and monitoring) at: http://www.compbio.dundee.ac.uk/jpred4/help.shtml

Creating individual inputs in the directory: `test_seqs.fa_dir`..............

**COMPLETED.**

14/05/15 13:59:29 ~/test > ./prepareInputs.csh test_seqs.fa

14/05/15 13:59:29 ~/test > ./prepareInputs.csh test_seqs.fa

14/05/15 13:59:36 ~/test > ls -ltrh test_seqs.fa_dir/

```
total 0
-rw-r--r-- 1 adrozdzetskiy lsd 68 May 14 13:59 IP_057216.1.OR4F5_1_1.fasta
-rw-r--r-- 1 adrozdzetskiy lsd 54 May 14 13:59 IP_057218.1.OR4F29_2.fasta
-rw-r--r-- 1 adrozdzetskiy lsd 52 May 14 13:59 IP_057219.1.OR4F29_3.fasta
-rw-r--r-- 1 adrozdzetskiy lsd 96 May 14 13:59 IP_057220.1.SAMD11_1_4.fasta
-rw-r--r-- 1 adrozdzetskiy lsd 88 May 14 13:59 IP_057221.1.SAMD11_1_5.fasta
-rw-r--r-- 1 adrozdzetskiy lsd 71 May 14 13:59 IP_057222.1.SAMD11_1_6.fasta
```

Check individual sequence FASTA files created.

**Note:** first word from the FASTA annotations (previous page) were used for individual file names.
Run automatic mass submission, status check and results retrieval script

14/05/15 14:15:07 ~/test > ./massSubmitScheduler.csh test_seqs.fa_dir

JPred REST API: mass submission and scheduling script (v.1)
Author: Dr. Alexey Drozdetskiy
Email: a.drozdetskiy@dunder.ac.uk
The Barton Group, Division of Computational Biology
College of Life Sciences, University of Dundee
www.compbio.dundee.ac.uk

Please contact me with suggestions to improve/fix client/server part of the JPred API and the mass submission/scheduling scripts.

Usage:

  ./massSubmitScheduler.csh directoryname

where directoryname - is the name of the local directory with single sequence FASTA format files you would want to submit as individual jobs.

NOTE: multiple controlled individual sequence submission through the JPred REST API is a preferred way over single batch submission. For details - please see tutorials on mass submission (and monitoring) at: http://www.compbio.dundee.ac.uk/jpred4/help.shtml

Script will start to submit jobs until reaching set limit (for test below we used limit=3).

After that automatically the script will check every submitted job status, once a job is complete, its results will be retrieved, and a new job would be submitted in place until all jobs done. All automatically.
Example of more logging output

n jobs running now: 3
Going to check status of jobid: jp_my9zoqU name: IP_057216.1.OR4F5_1_1
Going to check status of jobid: jp_oDV50IH name: IP_057218.1.OR4F29_2
Going to check status of jobid: jp_J7vdjNN name: IP_057220.1.SAMD11_1_4
Jod jp_J7vdjNN finished. Getting results, creating job directory and moving there all the relevant files.
n jobs running now: 2
Input file: test_seqs.fa_dir/IP_057221.1.SAMD11_1_5.fasta submitting...
n jobs running now: 3
Going to check status of jobid: jp_my9zoqU name: IP_057216.1.OR4F5_1_1
Going to check status of jobid: jp_oDV50IH name: IP_057218.1.OR4F29_2
Going to check status of jobid: jp_aSpX_T name: IP_057221.1.SAMD11_1_5
n jobs running now: 3
Going to check status of jobid: jp_my9zoqU name: IP_057216.1.OR4F5_1_1
Going to check status of jobid: jp_oDV50IH name: IP_057218.1.OR4F29_2
Jod jp_oDV50IH finished. Getting results, creating job directory and moving there all the relevant files.
Going to check status of jobid: jp_aSpX_T name: IP_057221.1.SAMD11_1_5
n jobs running now: 2
Input file: test_seqs.fa_dir/IP_057222.1.SAMD11_1_6.fasta submitting...
n jobs running now: 3
Results, output

On 14/05/15 at 15:07:43, the command `ls -ltrh` was executed in the directory `~/test`.

The output shows a list of directories and files, including:

- `jpredapi`
- `test_seqs.fa`
- `prepareInputs.csh`
- `jpredMassSubmitSchedule.tar.gz`
- `test_seqs.fa_dir_error`
- `test_seqs.fa_dir`
- `test_seqs.fa_dir_output`
- `massSubmitScheduler.csh`

These directories were automatically created for successful output, errors...

A successful jobs output example includes:

* `*.fasta`
* `*.jnet`

These files contain a summary of the JPred/JNet predictions (as well as a copy of the input FILES files). See [http://www.compbio.dundee.ac.uk/jpred4/help.shtml](http://www.compbio.dundee.ac.uk/jpred4/help.shtml) for more details, tutorials on JPred results.

Automatically created directories: for successful output, errors...
More details…

• There would be more files created during submission, status checks and results retrieval. All of them are technical and could be ignored (these files might be useful whenever you trying to investigate a problem with job submissions yourself or communicating to JPred team).