[Biopython] Biopython Digest, Vol 124, Issue 9

Dan dan837446 at gmail.com
Thu Apr 11 16:51:13 EDT 2013

- Previous message: [Biopython] Request from help
- Next message: [Biopython] BioPython now available on PiCloud by default
- **Messages sorted by:** [date] [thread] [subject] [author]

This is peripherally relevant to the question, I asked Tao Tao of NCBI user services about general guidelines for remote blast, and got this response:

"In general, the key is to reduce the hits to BLAST server:
At the search step, DO NOT submit searches that contain only single sequence! You need to batch the query and submit a set in a single search request.
At the result polling step, you should reduce the result checking by spacing them out, and start checking for results after a delay (a few minutes).
The XML result for batch queries is a bit peculiar each query is wrapped around <Iteration> tag
You are better off leaving the other conditions default and post-process it to get the top hits"

Also it's best to search between 9PM and 5AM Eastern Standard time and at weekends.
Personally I seem to encounter glitches using batches above 100 but it's so specific to your particular workplace that I'm not sure if that's a good guideline.

On Fri, Apr 12, 2013 at 4:00 AM, <biopython-request at lists.open-bio.org>wrote:

> *Send Biopython mailing list submissions to*
> *biopython at lists.open-bio.org*
> *To subscribe or unsubscribe via the World Wide Web, visit*
> *http://lists.open-bio.org/mailman/listinfo/biopython*
> *or, via email, send a message with subject or body 'help' to*
> *biopython-request at lists.open-bio.org*
> *You can reach the person managing the list at*
> *biopython-owner at lists.open-bio.org*
> *When replying, please edit your Subject line so it is more specific*
> *than "Re: Contents of Biopython digest..."*
> *Today's Topics:*
> *1. query upper limit for NCBIWWW.qblast? (Matthias Schade)*
> *2. Re: query upper limit for NCBIWWW.qblast? (Peter Cock)*
> ---
> *Message: 1*
> *Date: Thu, 11 Apr 2013 11:20:31 +0200*
> *From: Matthias Schade <matthiasschade.de at googlemail.com>*
> *Subject: [Biopython] query upper limit for NCBIWWW.qblast?*

> To: biopython at lists.open-bio.org
> Message-ID: <5166805F.8060603 at googlemail.com>
> Content-Type: text/plain; charset=ISO-8859-15; format=flowed
>
> Hello everyone,
>
> is there an upper limit to how many sequences I can query via
> NCBIWWW.qblast at once?
>
> Sending up to 150 sequences each of 24mer length in a single string
> everything works fine. But now, I have tried the same for a string
> containing about 900 sequences. On good times, it takes the NCBI-server
> about 5min to send an answer. I save the answer and later open and parse
> the file by other functions in my code. However, even though I have
> queried the same 900 sequences, the resulting output-file varies in
> length (10 MB<x<20MB) and always at least misses the correct
> termination-tag in "<\BlastOutput>" or even misses more (this does not
> happen why querying 150 sequences or less).
>
> I would guess once the server has started sending its answers, there
> might only be a limited time NCBIWWW.qblast waits for follow up packets
> ... and thus depending on the current server-load, the
> NCBIWWW.qblast-function simply decides to terminate waiting for
> incomming data after some time, resulting in my blast-output-files to
> vary in length. Could anyone correct or verify this long-fetched
> hypothesis?
>
> My core-lines are:
>
> orgn='Mus Musculus' #on anything else
> result = NCBIWWW.qblast("blastn", "nt", fasta_seq_string, expect=100,
> entrez_query=str(orgn+"[orgn]"))
> save_file = open ('myblast_result.xml',"w")
> save_file.write(result.read())
>
> Best regards,
> Matthias
>
>
> ------------------------------
>
> Message: 2
> Date: Thu, 11 Apr 2013 10:43:44 +0100
> From: Peter Cock <p.j.a.cock at googlemail.com>
> Subject: Re: [Biopython] query upper limit for NCBIWWW.qblast?
> To: Matthias Schade <matthiasschade.de at googlemail.com>
> Cc: biopython at lists.open-bio.org
> Message-ID:
>   <CAKVJ-_6y_q8e=EV5+1vCCeRY5c8z-brOsyHWW960dG0bX=
> ZYEg at mail.gmail.com>
> Content-Type: text/plain; charset=ISO-8859-1
>
> On Thu, Apr 11, 2013 at 10:20 AM, Matthias Schade
> <matthiasschade.de at googlemail.com> wrote:
> > Hello everyone,
> >
> > is there an upper limit to how many sequences I can query via
> NCBIWWW.qblast
> > at once?
>
> There are sometimes limits on the URL length, especially if going via
> firewalls and proxies, so that may be one factor.
>
> At the NCBI end, I'm not sure what limits they impose on this:
> http://www.ncbi.nlm.nih.gov/BLAST/Doc/urlapi.html

> > Sending up to 150 sequences each of 24mer length in a single string
> > everything works fine. But now, I have tried the same for a string
> > containing about 900 sequences. On good times, it takes the NCBI-server
> > about 5min to send an answer. I save the answer and later open and parse
> the
> > file by other functions in my code. However, even though I have queried
> the
> > same 900 sequences, the resulting output-file varies in length (10
> > MB<x<20MB) and always at least misses the correct termination-tag in
> > "<\BlastOutput>" or even misses more (this does not happen why querying
> 150
> > sequences or less).
> >
> > I would guess once the server has started sending its answers, there
> might
> > only be a limited time NCBIWWW.qblast waits for follow up packets ... and
> > thus depending on the current server-load, the NCBIWWW.qblast-function
> > simply decides to terminate waiting for incomming data after some time,
> > resulting in my blast-output-files to vary in length. Could anyone
> correct
> > or verify this long-fetched hypothesis?
> >
> > My core-lines are:
> >
> > orgn='Mus Musculus' #on anything else
> > result = NCBIWWW.qblast("blastn", "nt", fasta_seq_string, expect=100,
> > entrez_query=str(orgn+"[orgn]"))
> > save_file = open ('myblast_result.xml',"w")
> > save_file.write(result.read())
> >
> > Best regards,
> > Matthias
>
> I think you've reach the scale where it would be better to run blastn
> locally - ideally on a cluster if you have access to one. You can
> download the whole NT database from here - most departments
> running BLAST with their own Linux servers will have a central copy
> which is kept automatically up to date:
> ftp://ftp.ncbi.nlm.nih.gov/blast/db/
>
> If you don't have those kinds of resources, then you can even
> run BLAST on your own Windows machine - although I'm not
> sure how much RAM would be recommended for the NT
> database which is pretty big.
>
> Regards,
>
> Peter
>
>
> ------------------------------
>
> _____
> Biopython mailing list - Biopython at lists.open-bio.org
> http://lists.open-bio.org/mailman/listinfo/biopython
>
>
> End of Biopython Digest, Vol 124, Issue 9
> **
>

- Previous message: [Biopython] Request from help

- Next message: [Biopython] BioPython now available on PiCloud by default
- **Messages sorted by:** [date] [thread] [subject] [author]

More information about the Biopython mailing list