Brief Manual of Gautomatch

****

Author: Dr. Kai Zhang, MRC Laboratory of Molecular Biology

Contact: kzhang@mrc-lmb.cam.ac.uk

Description: Gautomatch is a GPU accelerated program for accurate, fast, flexible and fully automatic particle picking from cryo-EM micrographs with or without templates.

****

Features :

- **Fast:** typically, **1.5~2.0s** with 15 templates, using a good GPU (e.g. GTX 980, Titan X);
- **Fully automatic** with simple command on entire data sets;
- **Convenient** and easy to use;
- **Flexible**: with or without template, suitable for both basic or advanced users;
- **Compatible** with Relion/EMAN;
- **Background correction:** automatic correct the gradient background that affects the picking;
- **Rejection of ice/carbon:** automatically detect non-particle areas and reject them;
- **Post-optimization:** scripts available to re-filter the coordinates after picking within seconds
- **Accuracy:** the users' satisfaction is the only 'gold standard' criterion;

Requirement :

--> CentOS/Redhat Linux x86_64 (might be problems on Ubuntu or SUSE)

--> Any one of the libraries from **CUDA 5.0** to **7.5,** but do use the right version of **Gautomatch** according to your CUDA version

--> GPU architecture (Compute Capability) >= SM 2.0 or SM 3.0 Find

the web more details about Cuda version, GPU Compute Capability:

https://en.wikipedia.org/wiki/CUDA

(for lower architecture, <sm2.0, please contact me)

****

Program Download :

http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/Gautomatch_v0.50_and_examples.tar.gz

(only pre-compiled binary available now; source code to be released soon when other related projects are done; Several examples and suggested commands included)

Usage:

Gautomatch [options] <micrographs>

Basic options: default values, description:

--apixM 1.34 Pixel size of the micrograph, in Ångstrom

--diameter 400 Particle diameter, in Ångstrom;

--T NONE Particle picking templates in 2D MRC stack; auto-generated if not provided; IMPORTANT: read the usage of option **--dont_invertT** for more information about templates

--apixT 1.34 Pixel size of the templates, in Ångstrom

Additional options(not suggested, only try to optimize in difficult cases), default values and description:

--ang_step 5 Angular step size for picking, invalid for auto-templates

--speed 2 Speed level {0,1,2,3,4}, the bigger the faster, but less accurate. However,

Suggested 2 for >1 MDa complex; 1 for <500kD complex; 1 or 2 for 500~1000 kD; 0 not suggested normally, because the 'accuracy' is simply fitting noise, unless for special noise-free micrographs; use 3 for huge virus, but 2 still preferred; probably do not use 4 at all, not accurate in general.

--boxsize 128 Box size in pixel, **NOT** in Ångstrom; By default a suggested value will be automatically calculated by --diameter and –apixM. It will use 1.3X diameter of your particle (by -diameter option). So don't worry about if not set.

--max_dist 300 Minimum distance between particles in Ångstrom; 0.9~1.1X diameter; can be 0.3~0.5 for filament-like. Don't be confused about the word --max_dist. This was a spelling error. It will be changed in future.

--cc_cutoff 0.1 Cross-correlation cutoff, 0.2~0.4 normally; Try to select several typical micrographs to optimize this value. Alternatively, it will be even faster if you use a small value, e.g. 0.1, first and then use 'box_filter.com' or 'box_filter2rl.com' to filter the box files afterwards. Script could be obtained here: http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/Gautomatch_v0.50/scripts/

Just run ./box_filter2rl.com Then it will tell you how to use.

Ice, contamination, aggregation, carbon edge, sharp gold/metal particles related options.

Carbon Edges:

--lsigma_cutoff 1.2 Local sigma cutoff (relative value), 1.2~1.5 should be a good range; normally a value >1.2 will be ice, protein aggregation or contamination. Try to decrease or increase it upon the '_rejected.box' file. This option is designed to get rid of sharp carbon/ice edges or sharp metal particles.

--lsigma_D 100 Diameter for estimation of local sigma, in Ångstrom; usually this diameter could be 0.5~2.0X of your particle diameter according to several factors. A diameter around 100~400 works best in most cases. Try to decrease or increase it upon the '_rejected.box' file. Using bigger --lsigma_D, normally you should decrease the --lsigma_cutoff. For smaller and shaper high density contamination/ice/metal particles, you could use a smaller --lsigma_D and bigger --lsigma_cutoff.

Ice/Contamination:

--lave_min -1.0 Local average density cutoff (relative value), any pixel value below that will be considered as ice/aggregation/carbon etc. For 'black' cryoEM micrograph, set this to very small value e.g. -10.0 will not reject any 'black' dots in general. This option mainly rejects the central parts of the ice, carbon etc. which normally have lower density than the particles. Increase the value from -1.0 to -0.5 will reject more ice area, but may also reject your particles. Check the '_rejected.box' to check if the parameter is fine or not. Decrease the value from -1.0 to -1.5 will reject less ice area, but may mistake this area as your particles. Check the '_rejected.box' to optimize the best parameter for most of your micrographs.

--lave_max 1.0 Local average density cutoff (relative value), any pixel value above that will be considered as ice/aggregation/carbon etc (for contrast inverted 'white' micrograph or negative stain with big write blobs etc.). Normally, it is not useful for micrographs with 'black' particles, but might be helpful to get rid of 'hot' area. For negative stain micrograph, if it rejects most of the true particles, just use very big value, like 10.0, so that it will not reject anything.

--lave_D 400 Diameter for estimation of local average density, 0.5~2.0X particle diameter suggested; However, if you have 'sharp'/'small' ice or any 'dark'/'bright' dots, use a smaller value will be much better to get rid of these areas. It is quite similar to --lsigma_D, if you use bigger --lave_D, usually it is suggested to use smaller value of --lave_max(e.g. from 1.0 to 0.8) or bigger value of --lave_min (e.g. from -1.0 to -0.5) according the purpose.

--lp 30 Low-pass filter to increase the contrast of raw micrographs, suggested range 20~50Å. For bigger particles, use bigger low-pass will work better, smaller particles, use a bit smaller low pass, but <20 Å low pass is not suggested because there is no usable information due to CTF. You could use Gctf to do a phase flip first and use higher resolution for picking. But this is NOT suggested in general! First, it risking in the so-called 'Einstein noise' to pick more background.

Second, if you cannot properly pick the particle using lower resolution. That means the contrast is in big problem and the reconstruction will be not reliable in general.

--hp 1000 High-pass filter to get rid of the global background of raw micrographs, suggested range 200~2000Å. Don't worry about this option. Gautomatch has its approach to estimate and get rid of the background within the program anyway.

I/O options (use these for initial learning and diagnosis, no need for for the final jobs on whole-datasets):

--write_ccmax Specify to write out cross-correlation files

--write_pf_mic Specify to write out phase-flipped(pf) micrographs(mic)

--write_lave_mic Specify to write out estimated background(bg) of the micrographs(mic)

--write_bgfree_mic Specify to write out background subtracted (bgfree) micrographs(mic)

--write_lsigma_mic Specify to write out local sigma (lsigma) micrographs(mic)

--write_mic_mask Specify to write out the auto-detected mask (ice, contamination, aggregation, carbon edge etc.) by --lsigma_cutoff or --lave_max or --lave_max

--do_unfinished Specify to autopick the unfinished micrographs

--dont_invertT Whether to invert template contrast. VERY IMPORTANT!!! By default, the program will invert the 'white' templates to 'black' before picking. Specify this option to avoid contrast inversion if the micrographs and templates have the same contrast

--extract_raw Specify to extract particle from raw micrograph

--extract_pf Specify to extract particle from phase-flipped micrograph; will write a new stack and will not overwrite raw particle stack

--gid 0 GPU id, normally it's 0, use gpu_info to get information of all available GPUs.

Examples:

Gautomatch --apixM 1.58 --diameter 300 Micrographs/Falcon*.mrc (auto-generated templates for 'black' cryoEM micrograph)

Gautomatch --apixM 1.08 --diameter 300 --T templates.mrcs --apixT 2.16 Micrographs/Falcon*.mrc

(for 'white' templates and 'black' micrograph)

Gautomatch –apixM 1.08 --diameter 300 --T templates.mrcs --apixT 3.2 -dont_invertT Micrographs/Falcon*.mrc

(for 'black' templates and 'black' micrograph)

Gautomatch --apixM 1.08 --diameter 300 --boxsize 360 --write_bgfree_mic -write_lsigma_mic --extract_raw --write_ccmax Micrographs/Falcon*.mrc

(suggested for manual diagnosis using the different types of output micrographs)

More Examples and suggested commands:

http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/Gautomatch_v0.50/examples/

Useful scripts:

http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/Gautomatch_v0.50/scripts/

General Tips:

A better way is you split the entire datasets into several groups (e.g. 3-5) with similar appearances, and then optimize the parameters for each group.

It is suggested to run it outside the Micrographs/ directory to match typical Relion style.

There are several script that you can convert the .box file to reliant .star file. Using these script will be very helpful for the optimization of parameters, checking results, post-filtering the coordinates etc.

http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/Gautomatch_v0.50/scripts/

There are several I/O options that you can use to understand more about how Gautomatch works and also greatly help for troubleshooting or manual diagnosis and improvement of parameters.

Tips about CTF:

It is perfectly fine to use raw micrograph (before CTF correction) for particle picking since 30~50Å is sufficient to auto-pick the particles. Usually for the micrograph with defocus around 2-5um, the first 'zero-node' is around 20~30Å; So it is not useful to do CTF correction in general.

However, you can use 'Gctf' to automatically determine CTF and flip the phases before picking.

Fully CTF correction on micrograph or applying full CTF on templates is **NOT** suggested, because these operation is normally targeting for high resolution and performed in the last step during/after reconstruction.

Since particle picking is basically 'low-resolution' operation, higher resolution will only introduce more false picking and template-bias, known as the so-called 'Einstein noise' -------- A most risking thing in cryo-EM field for beginners !!!

Tips about '--dont_invertT' option:

This option is very important! By default, the program will invert the 'white' templates to 'black' before picking. This is because our cryoEM micrographs are usually 'black' and 2D averages are 'white'.

Specify '--dont_invertT' to avoid the contrast conversion if your micrographs and templates have the same contrast (either black or white).

Note that the auto-generated templates is ALWAYS 'white' Guassian blob, so for 'black' cryoEM, you should use default.

For 'negative stain' +'auto-templates', you should specify '--dont_invertT' so that the auto-generated templates and micrographs are both 'white'.

Tips about '--speed' option:

Suggested 2 for >1 MDa complex; 1 for <500 kD complex; 1 or 2 for 500~1000 kD; 0 not suggested, because the 'accuracy' is simply fitting noise, unless for special noise-free micrographs; use 3 for huge virus, but 2 still preferred; probably do not use 4 at all, not accurate in general.

In theory, a smaller value for --speed will generate a more accurate picking. But this is NOT true, because the meaning of ‘accuracy’ always risk in the situation of ‘higher noise’. So actually, --speed 2 works best in most cases because and --speed 1 might be better for smaller particles.