THe BuScr, ver. 1.07a
An interactive Triple-Helical collagen Building Script

Freely downloadable at structbio.biochem.dal.ca/jrainey

© Jan K. Rainey, 2003-2021.

Contact information:
Dr. Jan K. Rainey
Department of Biochemistry & Molecular Biology
Dalhousie University
Sir Charles Tupper Medical Building
5850 College Street
Halifax, Nova Scotia B3H 4R2
Canada
E-mail: jan.rainey@dal.ca

Licensed under the GNU General Public License. See accompanying license.txt for terms and conditions.

Key references

This script is presented in:
J.K. Rainey and M.C. Goh. (2004) An interactive triple-helical collagen builder. *Bioinformatics*, **20**, 2458-2459.
Please reference this publication if you use the script in your work.

The integral reference describing the statistical basis of the backbone building process is:
J.K. Rainey and M.C. Goh. (2002) A statistically derived parameterization for the collagen triple-helix. *Protein Science*, **11**, 2748-2754.

Melting temperatures for Gly-X-Y triplets are taken from:
A.V. Persikov, J.A. Ramshaw, A. Kirkpatrick, and B. Brodsky. (2000) Amino acid propensities for the collagen triple-helix. *Biochemistry*, **39**, 14960-14967.
A.V. Persikov, J.A. Ramshaw, A. Kirkpatrick, and B. Brodsky. (2002) Peptide investigations of pairwise interactions in the collagen triple-helix. *Journal of Molecular Biology*, **316**, 385-394.

The parameter set employed for prediction of terminal side-chain atoms and coordinate framework used to predict all-atom Hyp and Pro side-chains is developed in detail in:
J.K. Rainey and M.C. Goh. (2004) Statistically based reduced representation of amino acid side-chains. *Journal of Chemical Information and Computer Sciences,* **44**, 817-830.

Installing and Running THe BuScr

First things first – THe BuScr, over its lifetime to date, has been written in Tcl/Tk 8.4-2-8.5.4. There may be problems running the script under versions of Tcl/Tk previous to 8.5.4, therefore upgrading to (or starting with, for those have not had Tcl/Tk experience) the most recent version is probably a good idea if you are not already at 8.5.4 or above. (Go to www.tcl.tk for the latest, free, download.)

Upon gunzipping and tar extracting THe BuScr, you will have a new directory called THeBuScr.1.07 containing all necessary files to run the script. The following files will appear:

THeBuScr.1.07a.tcl – the actual Tcl/Tk script

HypScwrl3.tcl – tcl script to correctly amalgamate output of SCWRL with output of THe BuScr (see point (7) below)

THpropensities.dat – text file containing triplet propensity values for triple-helix

THparams_IR.dat – parameter set for imino rich helix type

THparams_AR.dat – parameter set for amino rich helix type

THparams_nonsmooth.dat – parameter set for transition between IR & AR occurring over a single peptide bond

THparams_smooth.dat – parameter set for transition between IR & AR occurring over three residues in each chain

THparams_chi1.dat – statistical χ_1 values for X & Y positions

ATlist.dat – list of terminal atoms for each residue handled by THeBuScr

ATdata/ - folder containing text file format terminal atom configuration data

THeBuScr.1.07a.pdf – Adobe PDF format manual (this file)

license.txt – the ubiquitous license file

updates_requests.txt – update history and list of further requested upgrades

README – this section in text format

UNIX (or Mac OS X terminal/X11)

For the present, no environment variables are used to direct THe BuScr to a particular path. Therefore, you need to run the script (THeBuScr.1.07a.tcl should be executable at the command prompt) from the directory containing the three TH*.dat files. The default version of the script assumes that the wish shell of Tcl/Tk is in the path:

```
#!/bin/sh
# Next line a tcl comment \
exec wish -f "$0" ${1+"$@"}
```

If this is not the case, you can edit the script to reflect the path to your wish shell as follows:

```
#!/bin/sh
# Next line a tcl comment \
exec /pathtotcl.tk/wish -f "$0" ${1+"$@"}
```

The exec line is useful in the case where Tcl/Tk is buried in a path of length more than 32 characters. The " –f" is included in this 32 characters. To launch THe BuScr, you would need the following two commands at the terminal:

```
cd /path_to_source/THeBuScr.1.07a/
THeBuScr.1.07a.tcl
```

Macintosh

Recent versions of Tcl/Tk seem not to automatically run Tcl/Tk scripts using Wish from Finder. Rather than packaging THe BuScr as a "standalone" application, I've opted to keep it very transparent in terms of where files are located etc. to allow for easy customization by you, the user. There are two options for running THe BuScr on a Mac:

1) The less preferred, but easier: If you double click on "THeBuScr.1.07a.tcl" in the folder "THeBuScr.1.07a" and Tcl/Tk is correctly installed, the application Wish (the Tk interpreter) should launch. However, it will not [generally] automatically run the tcl/tk script you have double clicked on. You can use then the Source command in the File menu to navigate through your computers folder system to choose THeBuScr.1.07a.tcl to run. ***However***, this appears to not size windows correctly, so you will need to drag the THe BuScr Tk window so that it is large enough to see all buttons etc. (See screen capture below.)

2) The surefire way to get around Finder's issues and ensure correct operation of THe BuScr is to use either Terminal or X11 to launch THe BuScr. After launching either terminal or X11, you can run THe BuScr using exactly the same two lines at the command prompt that are shown above for Unix. (Again, assuming Tcl/Tk is correctly installed.)

Windows

Once you've installed Tcl/Tk and rebooted, THeBuScr.1.07a.tcl should be recognized in Windows Explorer as a Tcl/Tk script. You can double click on the file, then to run it under the wish interpreter. This will ensure that all necessary .dat files are also found, since wish will then start in the THeBuScr.1.07a. folder.

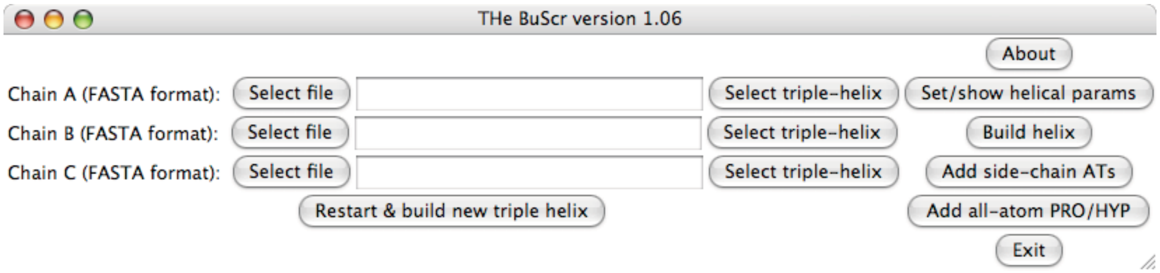
Building a triple-helical collagen molecule – Step by step

Important note: THe BuScr ver. 1.07a has been implemented assuming that your triple-helix architecture is as follows:

1) the three chains are either equal in length, or the first chain has one more triplet than the other two;
2) each chain starts with a Gly residue and ends with a Y residue.

These requirements are not set in stone, and improvements may be made in the future to allow more flexible handling of helices. For the present, however, you may encounter bugs if you try to use nonstandard architectures.

In general, THe BuScr will not let you proceed to a subsequent step if it does not have all required information. Therefore, many users may prefer to simply try running the script and refer to these detailed descriptions of each step when things aren't entirely clear. After running the script, you should see a window resembling the following, with various differences in appearance depending on your operating system:

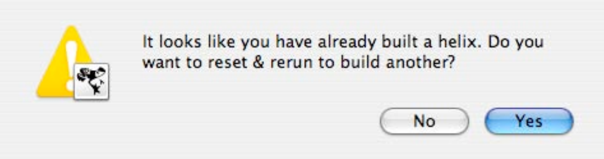


1) Specify FASTA sequence file(s) for each of the three chains of the triple-helix.

Note 1: *the Chain A, B and C you specify here need not be in a specific order to start with. THe BuScr will ask you if you would like to reorder them such that the chain with the largest number of triplets becomes Chain A for helical generation.* ***However, if you wish to explicitly specify the order, ensure that you select the appropriate chains: Chain A will start in the most N-terminal position, C in the most C-terminal position.***

Note 2: you can specify hydoxyproline using the one letter code O. For calculation of helical propensity, any P's in the Y-position will be automatically converted to O's. At PDB file generation time you have a choice to modify all P residues falling in the Y position to Hyp or to leave them as Pro.

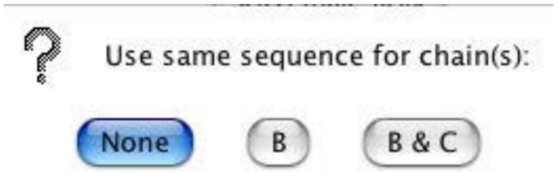
Note 3: If you want to build more than one triple-helical molecule in one session of THe BuScr it is HIGHLY RECOMMENDED that you click on "Restart & build new triple-helix" before building a subsequent model. The script will try to prevent this problem by presenting you with the following window:



It looks like you have already built a helix. Do you want to reset & rerun to build another?

However, the best way to prevent issues is to make sure you click on the "Restart" button when you want to build a second model.

If you wish to interactively select the file for each chain, simply click the "Select file" button beside the chain in question. (You may also enter a file name directly into the entry box between the "Select file" and "Select triple-helix" buttons.) If you interactively select Chain A, you will be asked:



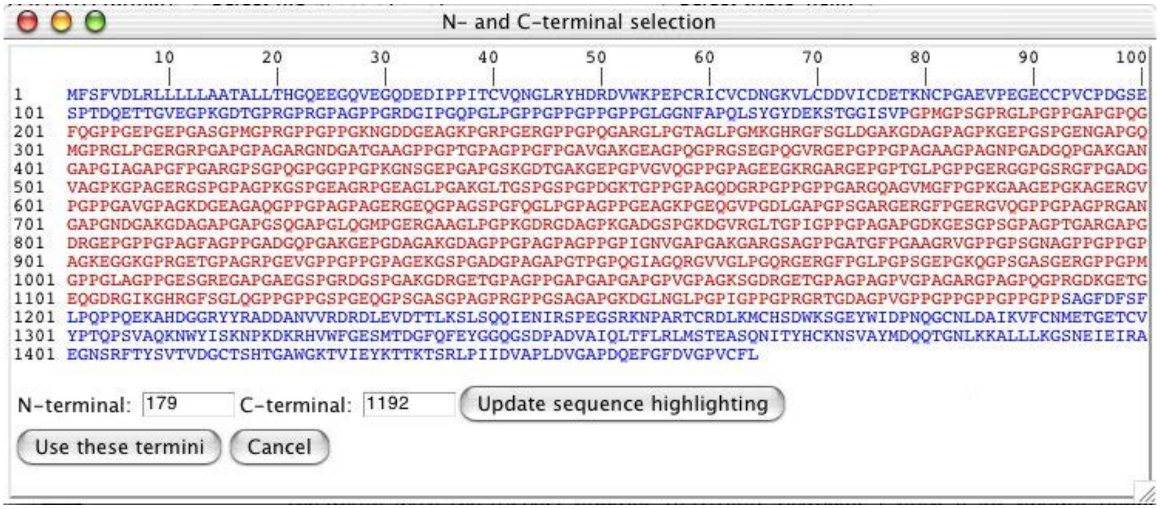
Use same sequence for chain(s):

In other words, if your triple helix is a homotrimer, where all three chains have the same primary sequence, you can simply select the "B&C" button to use the same FASTA file for all 3 chains. Conversely, if it's a heterotrimer select either "None" or "B", as appropriate.

Note: The default setting for any proline residues in Y positions is to treat them as hydroxyproline for helical propensity calculation. Please contact me if you'd like this to be changed.

2) Select the N- and C-terminal residues for chains.

Press the "Select triple-helix" button for Chain A. You should see a window that looks something like:



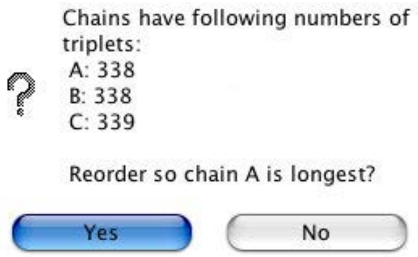
```
           10        20        30        40        50        60        70        80        90       100
1    MFSFVDLRLLLLLAATALLTHGQEEGQVEGQDEDIPPITCVQNGLRYHDRDVWKPEPCRICVCDNGKVLCDDVICDETKNCPGAEVPEGECCPVCPDGSE
101  SPTDQETTGVEGPKGDTGPRGPRGPAGPPGRDGIPGQPGLPGPPGPPGPPGPPGLGGNFAPQLSYGYDEKSTGGISVPGPMGPSGPRGLPGPPGAPGPQG
201  FQGPPGEPGEPGASGPMGPRGPPGPPGKNGDDGEAGKPGRPGERGPPGPQGARGLPGTAGLPGMKGHRGFSGLDGAKGDAGPAGPKGEPGSPGENGAPGQ
301  MGPRGLPGERGRPGAPGPAGARGNDGATGAAGPPGPTGPAGPPGFPGAVGAKGEAGPQGPRGSEGPQGVRGEPGPPGPAGAAGPAGNPGADGQPGAKGAN
401  GAPGIAGAPGFPGARGPSGPQGPGGPPGPKGNSGEPGAPGSKGDTGAKGEPGPVGVQGPPGPAGEEGKRGARGEPGPTGLPGPPGERGGPGSRGFPGADG
501  VAGPKGPAGERGSPGPAGPKGSPGEAGRPGEAGLPGAKGLTGSPGSPGPDGKTGPPGPAGQDGRPGPPGPPGARGQAGVMGFPGPKGAAGEPGKAGERGV
601  PGPPGAVGPAGKDGEAGAQGPPGPAGPAGERGEQGPAGSPGFQGLPGPAGPPGEAGKPGEQGVPGDLGAPGPSGARGERGFPGERGVQGPPGPAGPRGAN
701  GAPGNDGAKGDAGAPGAPGSQGAPGLQGMPGERGAAGLPGPKGDRGDAGPKGADGSPGKDGVRGLTGPIGPPGPAGAPGDKGESGPSGPAGPTGARGAPG
801  DRGEPGPPGPAGFAGPPGADGQPGAKGEPGDAGAKGDAGPPGPAGPAGPPGPIGNVGAPGAKGARGSAGPPGATGFPGAAGRVGPPGPSGNAGPPGPPGP
901  AGKEGGKGPRGETGPAGRPGEVGPPGPPGPAGEKGSPGADGPAGAPGTPGPQGIAGQRGVVGLPGQRGERGFPGLPGPSGEPGKQGPSGASGERGPPGPM
1001 GPPGLAGPPGESGREGAPGAEGSPGRDGSPGAKGDRGETGPAGPPGAPGAPGAPGPVGPAGKSGDRGETGPAGPAGPVGPAGARGPAGPQGPRGDKGETG
1101 EQGDRGIKGHRGFSGLQGPPGPPGSPGEQGPSGASGPAGPRGPPGSAGAPGKDGLNGLPGPIGPPGPRGRTGDAGPVGPPGPPGPPGPPGPPSAGFDFSF
1201 LPQPPQEKAHDGGRYYRADDANVVRDRDLEVDTTLKSLSQQIENIRSPEGSRKNPARTCRDLKMCHSDWKSGEYWIDPNQGCNLDAIKVFCNMETGETCV
1301 YPTQPSVAQKNWYISKNPKDKRHVWFGESMTDGFQFEYGGQGSDPADVAIQLTFLRLMSTEASQNITYHCKNSVAYMDQQTGNLKKALLLKGSNEIEIRA
1401 EGNSRFTYSVTVDGCTSHTGAWGKTVIEYKTTKTSRLPIIDVAPLDVGAPDQEFGFDVGPVCFL
```

N-terminal: 179 C-terminal: 1192

THe BuScr tries to guess the triple-helical region, but the "algorithm" used is a little crude. Therefore, you are forced to verify the N- and C-terminal residues whether you like it or not. The initially proposed triple-helical region is highlighted in red, while everything else is blue. If you like what you see, simply click "Use these termini". If not, you can update one or both termini and verify that your new sequence selections are what you expect with "Update sequence highlighting". If you click Cancel, no changes will be made to any previously selected termini. If you copied chain files after using the "File Select" dialog box, you will be given the option to use the same termini for each of these chains. ***Note that if you have modified the sequence file names for any of these copied files since using "File Select", THe BuScr is not going to flag this as a problem and will merrily assume that the file is the same. A good indicator that you don't have appropriate N- or C-terminal residues specified is a series of error messages in your Tcl script window (or xterm) saying "Non-Gly in G position - returning 0 propensity. Triplet: AAA" where AAA should be GXY for an appropriate triplet.***

3) Predict triple-helical propensity & interactively set helical parameters.

Once you specify all six N- and C-terminal residues, click on the "Set/show helical params" button. First, THe BuScr will tell you which N- and C-terminal residues you've selected for each chain, just to make sure you're happy with those selections. Next, you may be confronted with:



Chains have following numbers of triplets:
A: 338
B: 338
C: 339

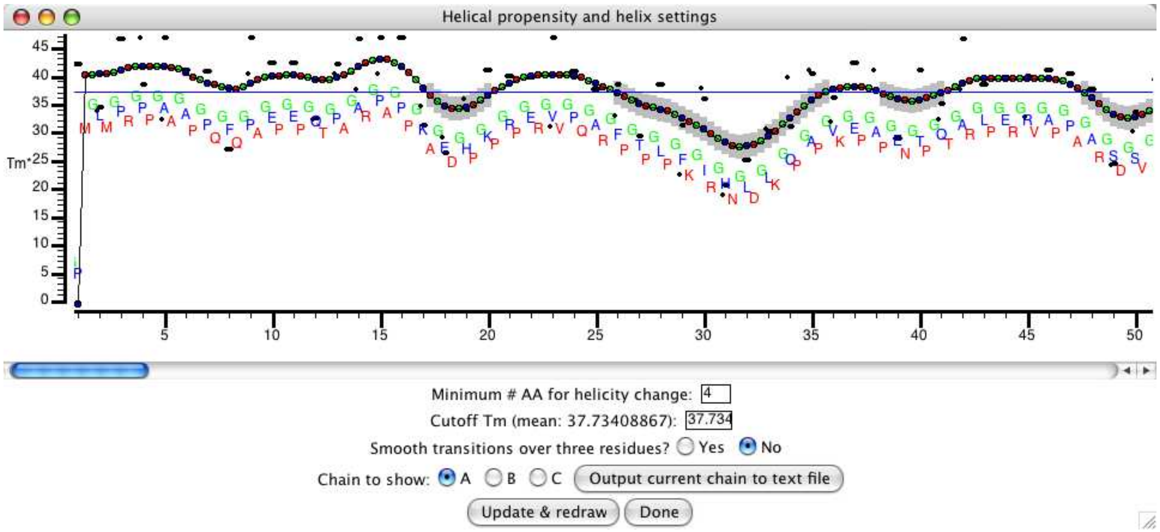
Reorder so chain A is longest?

A perhaps naive physical chemist such as myself assumes that the longest of the three chains should be in the A position to give it a chance to form a complete triple helix if it's only one triplet longer than the other two chains (as in type I collagen, for example) – therefore, THe BuScr will ask if you'd like to go ahead and make that longer chain into Chain A. The biochemical evidence to support or kill this assumption may not be available yet, however! If you are happy with the order of the chains that you initially specified, go ahead and click "No". Currently, THe BuScr will just go ahead and happily build one chain longer than the others as if it was continuing a normal triple-helix!

Now, helical propensities will be calculated. These are by default based upon melting temperature data from two papers by Persikov *et al.* given in the References section. You can change the manner of propensity calculations, update these values, add additional values etc. This is described in more detail later in the manual.

A sliding-window boxcar type average is carried out along the length of the helix in order to calculate predicted local melting temperatures (or other propensities). The exact details of this averaging scheme are provided later in the text.

Once the calculation is completed, you are presented with a window along the lines of:


Helical propensity and helix settings
Tm*
0
5
10
15
20
25
30
35
40
45
5
10
15
20
25
30
35
40
45
50
Output current chain to text file
Update & redraw
Done

Minimum # AA for helicity change: 4

Cutoff Tm (mean: 37.73408867): 37.734

Smooth transitions over three residues? ○ Yes ● No

Chain to show: ● A ○ B ○ C

While this window may appear a little bit busy, hopefully it will allow you sufficient flexibility to really examine helical propensity in a fairly intuitive way. Here is what it presents:

- The x-axis is triplet number, starting with 1. *The slider bar will let you scan your way along the entire length of the chain.*
- The coloured circles represent the predicted local melting temperature, T_m*, for the chain selected by the “Chain to show” radio button. Gly is green, X blue, Y red.
- The coloured letters are the one-letter code for each residue, falling directly underneath (or above when value is near 0) its T_m*.
- The black circles indicate the actual T_m for a given triplet. Chain A is to the left, Chain B in the middle, Chain C to the right. In the future, this may be made a little more elaborate – for the present, it’s just meant to aid in ensuring that you believe the calculated T_m*.
- The blue horizontal line shows the “Cutoff T_m” value. This defaults to the mean T_m: mean T_m is simply an arithmetic mean of the T_m for each triplet of each chain, not of the T_m* values.

- Grey boxes behind a given residue indicate that it has been flagged to be in a helix of lower stability. Namely, it is in stretch containing at least "Minimum # AA for helicity change" residues falling below "Cutoff T_m". By default, these will go to the "amino rich" helix parameters given in Rainey & Goh (2002); any residues without a grey box behind them will go to the "imino rich" parameters. These parameters may also be tweaked by the user as desired. See below for a description of such tweaking.

If you want to stray from the default minimum length of 4 residues for a change in helicity, modify the Cutoff T_m, and/or view a different chain, use the text entry boxes and radio buttons to do so. **Note that the "smooth transition over 3 residues" option is not currently recommended** – "smoothing" leads to a worse PROCHECK score than the default sharp transition over one peptide bond. In either case, the transition is somewhat "forced" into adopting the new helical conformation. See the THparams_nonsmooth.dat and _smooth.dat files for details. Following any change(s), you need to click on "Update helix settings & redraw" in order to see your modified settings take effect graphically. However, if you click "Done", any changes you have made to minimum length and Cutoff T_m will still be made – *i.e.* you do not need to visualize your changes if you don't want to.

Forcing helicity: currently, the only way to select helicity is using the T_m* values and some cutoff T_m. (Contact me if you'd like to be able to do this differently, such as load in a text file containing specified helicities for each residue – this would be a fairly straightforward modification.) However, it's very easy to force an entire helix to be IR or AR. For example, you could set the T_m cutoff to be 0°C, which would make all residues IR; or, make the T_m cutoff 50°C to force all residues to be AR.

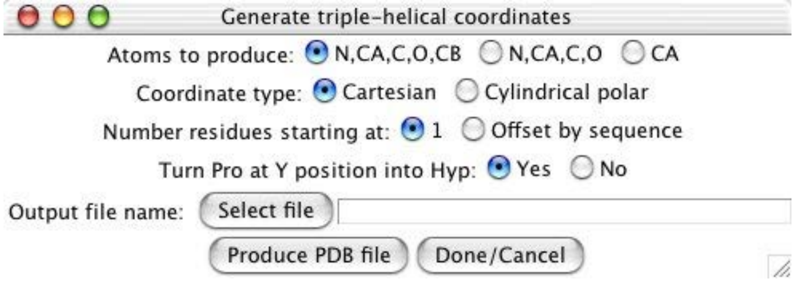
"Output current chain to text file" provides a useful file for further analysis/graphing of propensitites in external software. This is a space delimited file with the following 5 columns:

1) sequential number of residue (starting from 1)
2) three-letter code – Pro in Y is modified to Hyp
3) sequential number of triplet
4) predicted melting temperature of triplet
5) predicted local melting temperature of residue

Currently, output is to default filenames in the working directory – ChainA_prop.txt etc.

4) Build the triple-helical molecule.

After setting helical parameters and examining the propensities for your given helix, go ahead and press the "Build helix" button on the main THe BuScr window. This brings up yet another selection window:



Generate triple-helical coordinates

Atoms to produce: ● N,CA,C,O,CB ○ N,CA,C,O ○ CA

Coordinate type: ● Cartesian ○ Cylindrical polar

Number residues starting at: ● 1 ○ Offset by sequence

Turn Pro at Y position into Hyp: ● Yes ○ No

Output file name:

Here you have several options that are mostly self-explanatory. The "Number residues starting at" option allows you to have a numbering scheme in the PDB file for each chain which either starts at 1 or at the N-terminal number from your FASTA file (the "Offset by sequence" option). As with the FASTA file selection, you may use "Select file" to bring up an open file dialog box for ease of file entry.

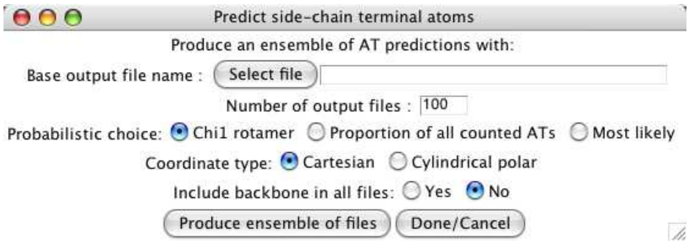
Once you are happy with your options, click on "Produce PDB file". You can go ahead and produce as many files as you want in as many formats as you want.

5) Predict side-chain terminal atoms.

THe BuScr is able to use a statistically derived library of reduced side-chain representations to position side-chain terminal atoms, or $\mathbf{A}^T$ groups. For full details of this reduced representation, see Rainey and Goh (2004). It requires N, C^α, and C' backbone atoms in Cartesian coordinates, but THe BuScr automatically produces a temporary pdb file if your most recently produced file does not have this information. As you will see in the *J Chem Inf Comput Sci* paper, the statistics are very accurate if χ_1 dihedral angles are used in prediction. Analysis was carried out using the exact same methodology as in Rainey and Goh (2002) for χ_1 dihedral angle preferences (Rainey 2003). Interestingly, the imino rich regions display exquisite specificity in χ_1 rotamer. The X position exhibits, practically exclusively, the g^+ rotamer with a filtered mean of 28.0° (σ 7.4°) based on 148 residues out of the initial data set of 168. In the Y position, the g^- rotamer is present in equal prevalence – the filtered data set of 178 out of 188 residues has a mean of –20.4° (σ 7.1°). In both of these filtering steps, the filtered mean is more than 1% away from that derived by four subsequent filters at $\mu \pm 2\sigma$. The means in the latter filtering case are actually closer to the ideal rotamer value: for X 29.5° (σ 3.8°) and for Y –21.2° (σ 5.2°). In either case, the χ_1 rotamers for X vs. Y significantly prefer g^+ and g^-, respectively. It would probably be unreasonable to expect that these should assume the ideal values of +60° and -60°, since the triple-helix is likely a deviant structure from the standard globular proteins in terms of side-chain rotamericity.

The amino rich region of 1BKV does not display such drastic specificity in χ_1 rotamer angles. A cursory examination of the non-Ala X positions shows a μ of -49.0° (σ 53.2°); for Y, μ is –26.9° (σ 69.6°). The large σ values are indicative of these μ values being not particularly representative. Three of the X positions are g^+, three are g^-, and one is t. The Y position is also spread between the three possible rotamers: four are g^-, three are t, and one is g^+. Without a larger non-imino triple-helical data set, it is meaningless to draw conclusions from these measurements.

When you click on the "Add side-chain ATs" button from the main THe BuScr window, you will be presented with the following window:



Predict side-chain terminal atoms

Produce an ensemble of AT predictions with:

Base output file name :

Number of output files : 100

Probabilistic choice: ● Chi1 rotamer ○ Proportion of all counted ATs ○ Most likely

Coordinate type: ● Cartesian ○ Cylindrical polar

Include backbone in all files: ○ Yes ● No

Here, you select a base output file name – to explain this, let's take `t1col` as an example base file name. The number of output files you specify will be the number of predictions carried out for each $\mathbf{A}^{\mathbf{T}}$ using the prediction method you specify: χ_1 rotamer (CHI1), proportion of all counted $\mathbf{A}^{\mathbf{T}}$'s (PC), or most likely position (ML); the codes in brackets correspond to the codes described in Rainey and Goh (2004). Note that for CHI1, PC will be used for all residues where no χ_1 rotamer is specified. The following PDB format files will then be produced with the specified coordinate type (using the example of "t1col" as base file name):

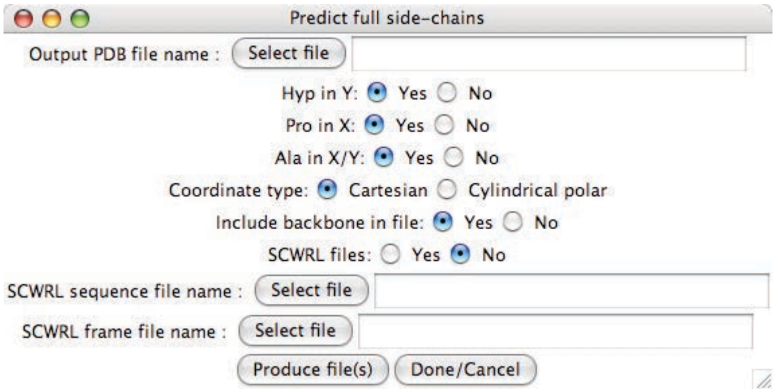
tlcol_bb.pdb – backbone coordinates

t1col_AT_#.pdb - **A^T** coordinates, with or without backbone coordinates as specified. Note: including backbone coordinates in every predicted file may end up producing a lot of MB of data for big ensembles, and big molecules.

6) Predict all-atom PRO in X-position and/or HYP in Y-position.

New as of version 1.06 is the ability to predict all-(non-H)-atom side chains for Pro in X and Hyp in Y. Note that Ala is also predicted in X and Y, if desired. **In order to use this functionality, you have to first specify helical parameters ("Set/show helical parameters" button) and then build a helix ("Build helix" button).** As usual, You will be reminded of this by THe BuScr if you forget.

If you click on "Add all-atom PRO/HYP" you will see:



Hyp in Y: ● Yes ○ No
Pro in X: ● Yes ○ No
Ala in X/Y: ● Yes ○ No
Coordinate type: ● Cartesian ○ Cylindrical polar
Include backbone in file: ● Yes ○ No
SCWRL files: ○ Yes ● No

The top six lines are fairly self-explanatory: Specify the output PDB file you want on the top line, and use the "Yes/No" selection radio buttons to specify the predictions you want to have carried out and included in the output PDB file. You can also specify coordinate types and inclusion or non-inclusion of backbone atoms in the PDB file.

7) Building a second triple-helical model with a different sequence

As mentioned above, you should reset the script in order to build a second model. Tcl/tk seems to have some issues otherwise. Simply click on the "Restart & build new triple-helix" button, at which point you will get the prompt:

Really reset and start again?

This is nothing to be scared of, just will start a new Wish session allowing you to robustly build a second (third, fourth, etc.) model.

8) Interfacing with SCWRL to predict remaining all-atom side-chains

Further all-atom predictions can be performed readily, although not quite automatically, using SCWRL 4.0 from the Dunbrack group (Krivov et al 2009). The sticky point is as follows: SCWRL doesn't incorporate amino acid types other than the standard 20 directly into its predicted structures. This necessitates a multi-step process, which I've tried to make as easy as possible short of full automation:

1) Use THe BuScr to predict backbone and any all-atom side-chains of the Hyp, Pro, and Ala set that you would like
2) Use SCWRL to predict remaining all-atom side-chains – when implemented using THe BuScr's output files, this will introduce Hyp/Pro/Ala side-chains as "road-blocks" to aid in correct prediction of the remainder of the side-chains
3) Use a supplementary tcl script provided with THe BuScr to amalgamate the outputs of SCWRL and THeBuScr, since the SCWRL output will have all Hyp residues replaced with Gly

To carry out steps 2 and 3, you first select "Yes" for SCWRL files in the "Predict full side-chains" window. Then, you need to specify names for your SCWRL sequence file (effectively tells SCWRL which residues to predict or leave alone) and SCWRL frame file (provides the positions of atoms for "nonstandard" Hyp side-chains). Click "Produce file(s)" and you should end up with 3 files produced.

Note: to run SCWRL, you need to use Cartesian coordinates and include the backbone in your PDB file.

Next up, you can leave THe BuScr. Let's say your output files were called t1_APO.pdb (for the PDB file with Ala, Pro and Hyp predicted), t1_scw.seq and t1_scw.frame (SCWRL sequence and frame files, respectively). You would then run SCWRL with the following options:

```
Scwrl4 –i t1_APO.pdb –u –s t1_scw.seq –f t1_scw.frame –o t1_scw.pdb
```

which will produce a PDB file t1_scw.pdb. (The –u option disables disulfide bond checking and is optional.) This may take some time to complete for a full-length collagen molecule, even on a fairly fast processor.

Finally, you can "amalgamate" your PDB files. I have provided a tcl script HypScwrl3.tcl (newly updated from HypScwrl2.tcl in THeBuScr 1.07) which will do this for you. There's no nice tk GUI yet – *if there is demand, I can incorporate this entire process right into the GUI of THeBuScr – e-mail me and let me know!* To run this tcl script, you need to first edit the script contents: the final line in the script calls a procedure "amalgamatePDB" and has three arguments: first, your THe BuScr PDB file (t1_APO.pdb in this example), second your SCWRL output file (t1_scw.pdb in the above example) and third, your desired amalgamated PDB file name. So, for the present example, you would edit this line to read:

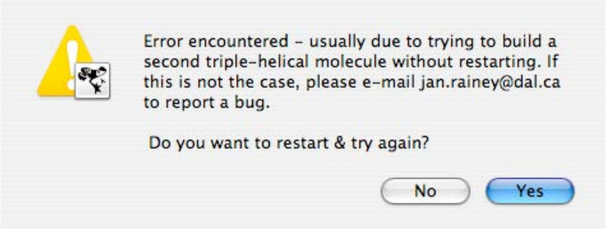
```
amalgamatePDB {t1_APO.pdb} {t1_scw.pdb} {t1col.pdb}
```

save the revised tcl script and then run the script either from the command prompt in a Unix type system or by double-clicking in Windows or Mac OSX. Documentation in the file reminds you about specifying path, if necessary.

As always, I am happy to provide further support by e-mail if you are having trouble with this process!!!

9) Error/bug?

If you encounter an error or bug in the script, you should see the following window:



Error encountered – usually due to trying to build a second triple-helical molecule without restarting. If this is not the case, please e-mail jan.rainey@dal.ca to report a bug.

Do you want to restart & try again?

As the message implies, most errors tend to be encountered if you've tried to build a second triple-helical molecule without resetting/rerunning the script. Just click "Yes", reload your sequences and try again. ***If the error persists, please do e-mail me! A file should also be produced in your THeBuScr.1.07a directory called "error_THeBuScr.out" containing details of the error that you've encountered.***

Further Customization Potential of THe BuScr

1) Editing propensities & using unnatural amino acids

To edit the propensity scale you need to modify the text file THpropensities.dat in the directory containing THe BuScr. Add triplet entries as required following the format “GXX ###” or edit the existing entries as desired. Note that lines starting with a semicolon are comments.

If you wish to introduce an unnatural amino acid, choose a one-letter code (this could theoretically be any character, not just a letter) other than the standard 21 one-letter codes (with O for hydroxyproline being the 21st). Add propensity values to THpropensities.dat as required. *No actual code changes are required,* although you will probably also want to edit THeBuScr.#.##.tcl itself as well to allow it to output your three-letter code of choice. The only change that should be required is to find “proc retthreelet { onelett }”, and to add your one-letter code along with a corresponding three-letter code following the format used for all other residues.

2) Modifying helical parameters.

The files THparams_AR.dat and THparams_IR.dat in the directory containing THe BuScr may be edited in order to modify helicity. These correspond to the “amino rich” and “imino rich” helical parameter sets given in Rainey & Goh (2002). Note that adding a third (or more) helical state is not trivial, but that updating the AR or IR parameter set is. Simply modify the desired atom radius, angle, or progression along the helical long-axis value (the x value in the .dat files) and this modification will be incorporated into any helices you build. Note that the carbonyl carbon parameters are labelled “CO” while the carbonyl oxygen parameters are “OC”, instead of the usual “C” and “O”. CA and CB refer as normal to C^{α} and C^{β}, respectively. Note that this may also require changing the correction factors in “THparams_[non]smooth.dat” to ensure that transitions lead to correct chain-to-chain spacing and angles.

Sliding window boxcar averaging methodology

The following section is updated from Jan K. Rainey's Ph.D. Thesis entitled "Collagen structure and preferential assembly explored by parallel microscopy and bioninformatics", at the University of Toronto, Toronto, Ontario, Canada, 2003.

As noted above, the effect of various amino acids substitutions on triple-helical stability has been extensively examined by B. Brodsky and co-workers using model host-guest peptides (Persikov et al 2000b and 2002). For triple-helical peptides, the circular dichroism (CD) ellipticity was monitored at 225 nm as the temperature was increased from 0°C to 80°C, From the ellipticity at a given temperature, the fraction folded at temperature T, F(T), was determined as:

$$F(T) = \frac{\theta(T) - \theta_U(T)}{\theta_N(T) - \theta_U(T)} \tag{A-1}$$

where $\theta(T)$ is the measured ellipticity and $\theta_U(T)$ and $\theta_N(T)$ are the ellipticities of the unfolded peptide and the folded triple helix, respectively. Although these latter baseline values are not quoted directly in either study, the melting temperature (T_m) appears to be simply the inflection point of a typical CD intensity vs. temperature melting curve. The residual ellipticity at 225 nm is significantly lower for the sample above T_m as compared to below T_m.

A host peptide of Ac-(Gly-Pro-Hyp)$_3$-**Gly-X-Y**-(Gly-Pro-Hyp)$_4$-Gly-Gly-CONH$_2$ containing Pro and Hyp in the boldface X and Y positions is used as a starting point in the Brodsky group studies. Peptides were synthesized with substitutions either at the X or Y position (Persikov *et al.*, 2000), or in both the X and Y positions simultaneously (Persikov *et al.*, 2002). This pair of studies provides T_m for 19 (Gly-X-Hyp), 19 (Gly-Pro-Y), and 41 (Gly-X-Y) guest triplets as compared to the Gly-Pro-Hyp host triplet. These results are given numerically in Table A-1. In the initial work by Persikov *et al.*, and a parallel paper describing analysis (Persikov et al 2000a), the melting temperatures for a given Gly-X-Y triplet were considered on a purely additive basis from those of the Gly-Pro-Y and Gly-X-Hyp triplets measured. The T_m for a given GXY triplet was then calculated as:

$$T_m(GXY) = T_m(GXO) + T_m(GPY) - T_m(GPO) \qquad (A\text{-}2)$$

However, an examination of the values listed in Table 8-1 shows that this is definitely not true experimentally. The most extreme demonstration is the triplet Gly-Lys-Asp. By additivity (eq. A-2), this would be expected to have a T_m of only 28.2°C; experimentally, it demonstrates a significantly higher T_m of 34.5°C (Chan et al 1997). Deviations such as this prompted the second study of Gly-X-Y triplets, where non-imino groups were present in both the X and Y positions (Persikov et al 2002). While fairly extensive thermodynamic analysis was carried out during these works, the most useful finding with respect to prediction of triple-helical structure is that the T_m values can be related directly to ΔG values, and therefore to the general propensity of a given triplet form a triple-helix. One algorithm for sequence specific stability calculation of a triple-helix is presented by Bächinger and Davis (1991) making use of earlier melting temperature data and examining tripeptides along the sequence. A slightly more elaborate scheme using a sliding window boxcar average is presented and used here.

Table A-1 – Melting temperatures (T_m) of various triple-helical peptides with indicated triplet as the fourth triplet from the N-terminal in an 8-triplet trimer. Determined by Persikov *et al.* (2000, 2002) using CD.

Triplet	T_m (°C)	Triplet	T_m (°C)	Triplet	T_m (°C)	Triplet	T_m (°C)
Gly-Pro-Hyp	47.3	Gly-Lys-Gln	38.9	Gly-Glu-Lys	35.0	Gly-Leu-Ala	31.2
Gly-Pro-Arg	47.2	Gly-Met-Hyp	38.6	Gly-Glu-Ala	34.6	Gly-Leu-Lys	31.1
Gly-Glu-Hyp	42.9	Gly-Ile-Hyp	38.4	Gly-Arg-Asp	34.5	Gly-Asp-Lys	30.9
Gly-Pro-Met	42.6	Gly-Asn-Hyp	38.3	Gly-Tyr-Hyp	34.3	Gly-Ala-Lys	30.8
Gly-Ala-Hyp	41.7	Gly-Ala-Arg	38.2	Gly-Pro-Asp	34.0	Gly-Arg-Ser	30.5
Gly-Pro-Ile	41.5	Gly-Ser-Hyp	38.0	Gly-Ile-Ala	33.9	Gly-Pro-Asn	30.3
Gly-Lys-Hyp	41.5	Gly-Pro-Cys	37.7	Gly-Arg-Glu	33.8	Gly-Pro-Tyr	30.2
Gly-Pro-Gln	41.3	Gly-Glu-Gln	37.7	Gly-Phe-Hyp	33.5	Gly-Glu-Asp	29.7
Gly-Pro-Ala	40.9	Gly-Asp-Arg	37.1	Gly-Gly-Hyp	33.2	Gly-Arg-Lys	29.5
Gly-Arg-Hyp	40.6	Gly-Pro-Lys	36.8	Gly-Ala-Asp	33.0	Gly-Glu-Asn	29.5
Gly-Gln-Hyp	40.4	Gly-His-Hyp	36.5	Gly-Ala-Ser	33.0	Gly-Pro-Phe	28.3
Gly-Glu-Arg	40.4	Gly-Thr-Hyp	36.2	Gly-Ala-Ala	32.9	Gly-Ala-Leu	27.8
Gly-Asp-Hyp	40.1	Gly-Cys-Hyp	36.1	Gly-Pro-Gly	32.7	Gly-Gly-Lys	26.9
Gly-Pro-Val	40.0	Gly-Glu-Thr	35.9	Gly-Gln-Lys	32.6	Gly-Leu-Leu	26.9
Gly-Pro-Glu	39.7	Gly-Lys-Asp	35.8	Gly-Val-Lys	32.5	Gly-Pro-Trp	26.1
Gly-Pro-Thr	39.7	Gly-Pro-His	35.7	Gly-Trp-Hyp	31.9	Gly-Gly-Ala	26.0
Gly-Gln-Arg	39.5	Gly-Leu-Gln	35.7	Gly-Pro-Leu	31.7	Gly-Gly-Leu	25.3
Gly-Lys-Arg	39.1	Gly-Glu-Val	35.3	Gly-Met-Lys	31.7	Gly-Phe-Ala	24.1
Gly-Leu-Hyp	39.0	Gly-Lys-Glu	35.3	Gly-Lys-Asn	31.7	Gly-Ala-Phe	21.9
Gly-Val-Hyp	38.9	Gly-Pro-Ser	35.0	Gly-Asp-Ala	31.6	Gly-Gly-Phe	19.7

The statistical parameterization developed by us (Rainey and Goh 2002) leads most directly to a two-state model for a general triple-helical peptide structure. While more elaborate models can be readily envisioned, this seems the most logical starting point for predicting a triple-helix structure in which only the fully folded helix is considered. Following from the two parameter sets provided, this will be assumed to be in one of:

1) IR - a more compact and tightly wound triple-helix, represented by the imino-rich parameter set; or,
2) AR - a more extended, less tightly would triple-helix, represented by the imino-deficient parameter set.

For simplicity in boxcar averaging, it will be assumed that the transition between two such regions occurs sharply.

We feel that the T_m values summarized in Table A-1 are the best data presently available to select between the helix types IR and AR. A sliding window three triplets in width is used to produce average T_m values over the entire length of the triple-helix. For the purposes of sliding window averaging, the 1/3 triplet stagger between residues is simplified to a one residue stagger to provide corresponding three triplet regions in the other two chains. (Although residues Gly→X, X→Y and Y→Gly are often simplified to be evenly spaced along the length dimension of the helix, our statistical parameters shown distinctly that each of these actually has a different displacement relative to the helical long-axis.) Given this one residue stagger assumption, the appropriate sliding window will be composed as shown in Figure A-1:

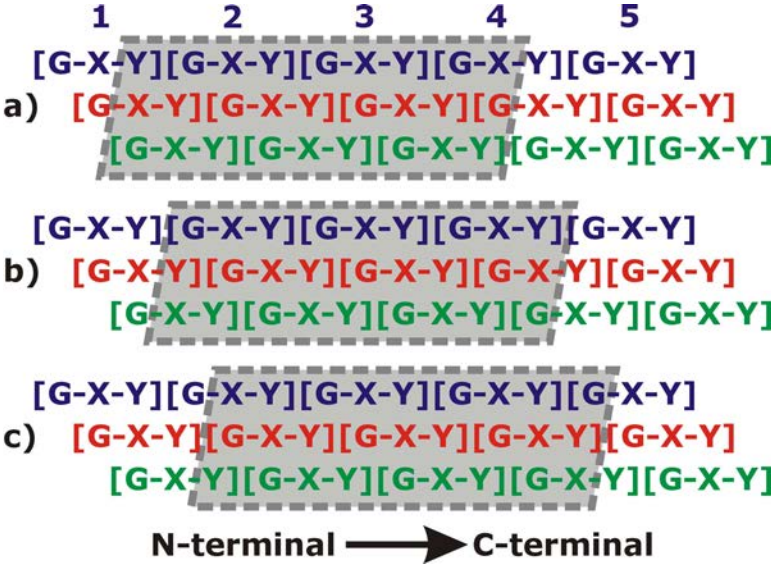

1
2
3
4
5
[G-X-Y][G-X-Y][G-X-Y][G-X-Y][G-X-Y]
a)
[G-X-Y][G-X-Y][G-X-Y][G-X-Y][G-X-Y]
[G-X-Y][G-X-Y][G-X-Y][G-X-Y][G-X-Y]
[G-X-Y][G-X-Y][G-X-Y][G-X-Y][G-X-Y]
b)
[G-X-Y][G-X-Y][G-X-Y][G-X-Y][G-X-Y]
[G-X-Y][G-X-Y][G-X-Y][G-X-Y][G-X-Y]
[G-X-Y][G-X-Y][G-X-Y][G-X-Y][G-X-Y]
c)
[G-X-Y][G-X-Y][G-X-Y][G-X-Y][G-X-Y]
[G-X-Y][G-X-Y][G-X-Y][G-X-Y][G-X-Y]
N-terminal
C-terminal

Figure A-1 – Schematic illustration of sliding window boxcar used for determination of local preference for triple-helicity. Five triplets from each of the three chains are illustrated, as numbered 1-5 on chain A (blue). (a) – (c) shows the three possible boxcar configurations, depending upon the current window. See the text for an exact description.

In considering Figure A-1, it is important to note that each triplet will have a defined T_m value from Table A-1, rather than each of the individual Gly-X-Y components. The T_m values determined by the Brodsky group are used as given in the above referenced papers; for any non-imino acid containing triplets not yet determined, pure additivity (eq. A-2) is used to approximate the T_m for the triplet. No attempts are made to include deviation from this ideal additivity, since Persikov *et al*. demonstrate that these deviations are not particularly predictable and may be in either direction.

As the window translates from the N-terminal to C-terminal, only three different configurations for the average are possible. These are depicted in Figure A-1(a)-(c). A further C-terminal translation by one residue will bring the configuration back to that in Figure A-1(a), centred about triplet 3 of Chain C instead of triplet 2. The exact contributions to each sliding window average are given in Table A-2. As with other studies, any prolines reported in Y positions from the primary sequence are assumed to be 4-hydroxyproline for this analysis.

Table A-2 – Three triplet wide sliding window average (Figure A-1) components for each of the three possible positions. Chain lettering convention is: A most N-terminally translated; C most C-terminal.

Window position in Figure A-1	Relative triplet number	Weighting of triplet T_m		
		Chain A	Chain B	Chain C
(a)	N-2	1/9	2/9	3/9
	N-1	3/9	3/9	3/9
	N	3/9	3/9	3/9
	N+1	2/9	1/9	0
(b)	N-2	0	1/9	2/9
	N-1	3/9	3/9	3/9
	N	3/9	3/9	3/9
	N+1	3/9	2/9	1/9
(c)	N-2	0	0	1/9
	N-1	2/9	3/9	3/9
	N	3/9	3/9	3/9
	N+1	3/9	3/9	2/9
	N+2	1/9	0	0

In order to provide a representative T_m, T_m*, for a given Gly, X, or Y residue, the sliding window averages are then boxcar averaged. This is done through the inclusion of every instance in which the given residue contributes to a sliding window average. The exact contributions from sliding window averages to the T_m* assigned to each residue are as given in Table A-3. This serves to very effectively weight the surrounding triplets decreasingly as the distance from the given residue increases. Furthermore, the Gly, X, and Y positions contribute differently in the weighting. For the Gly and Y positions, the N- or C-terminal triplets are represented to a greater degree, respectively, in the average. For the X position, both the N- and C-terminal triplet weightings are equivalent. Since each case is the sum of contributions from all three chains, every T_m* consists of one each of the Gly, X and Y configurations. Figures A-2 – A-4 show more clearly the total fractional contributions of the T_m of each surrounding triplet to a given T_m*.

Table A-3 – Assignment of T_m* values for each triplet position of each chain. Chain A is most N-terminally translated N-terminal residue; C most C-terminal.

Chain			**Calculation for T_m (representative)**[a]
A	**B**	**C**	
X_N	G_N	Y_{N-1}	$([(a) + (b) + (c)]_{N-1} + [(a) + (b) + (c)]_N + [(a) + (b) + (c)]_{N+1})$ x 1/9
Y_N	X_N	G_N	$([(b) + (c)]_{N-1} + [(a) + (b) + (c)]_N + [(a) + (b) + (c)]_{N+1} + [(a)]_{N+2})$ x 1/9
G_{N+1}	Y_N	X_N	$([(c)]_{N-1} + [(a) + (b) + (c)]_N + [(a) + (b) + (c)]_{N+1} + [(a)+(b)]_{N+2})$ x 1/9

a) (a), (b) and (c) refer to sliding window averages in Table A-2 for a given N.

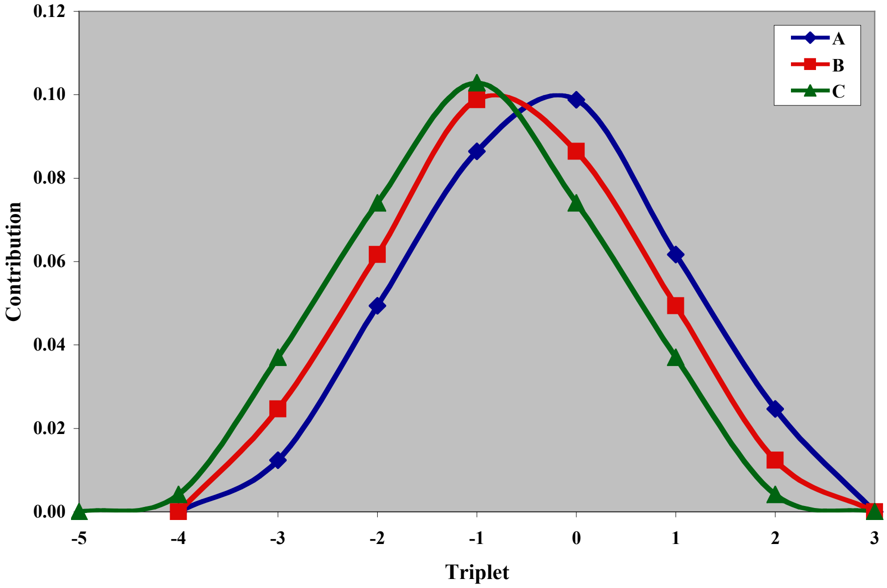


Figure A-2 – Weighting of contribution of the T_m for a given triplet of a given chain to the value of T_m* for residue G_0 of chain A, Y_{-1} of chain B and X_{-1} of chain C arising from the sliding window boxcar average employed in THe BuScr.

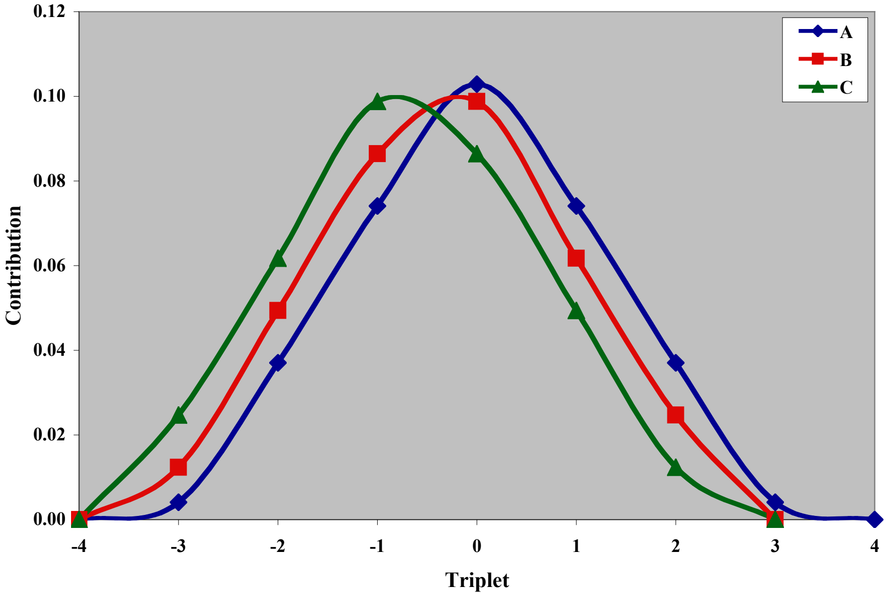


Figure A-3 – Weighting of contribution of the T_m for a given triplet of a given chain to the value of T_m* for residue X_0 of chain A, G_0 of chain B and Y_{-1} of chain C arising from the sliding window boxcar average employed in THe BuScr.

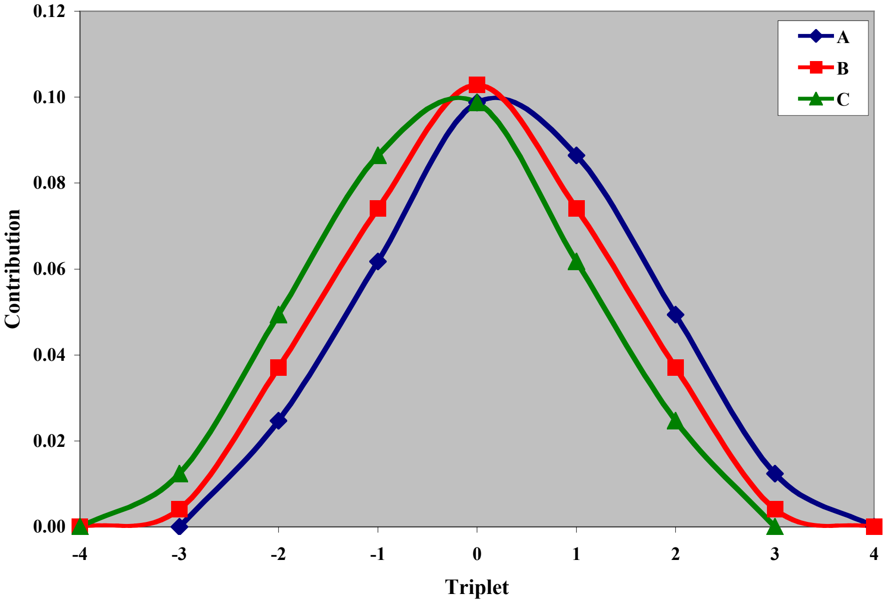


Figure A-4 – Weighting of contribution of the T_m for a given triplet of a given chain to the value of T_m* for residue Y_0 of chain A, X_0 of chain B and G_0 of chain C arising from the sliding window boxcar average employed in THe BuScr.

At the N- and C-terminal regions, the same style of averaging is employed, where the weighting functions shown in Figures A-2 – A-4 are moved closer and closer to the terminal. As the weighting function passes over the end of the helix, it loses its symmetric nature. This means that by the time it reaches a terminal set of aligned residues (such as Y_1 of A, X_1 of B and G_1 of C), the maximal contributors to the weighted T_m* will be those residues at the terminal. In this particular case, the weighting functions illustrated in Figure A-4 correspond exactly to the "triplet value" 0 of the figure for Y_1 of A, X_1 of B and G_1 of C. None of the negative triplets have meaning at this point, therefore all contributions arise at the N-terminal and in the subsequent 3 triplets. This potentially leads to an elevated prediction of helical stability at the terminal regions. However, for the present, we have decided to be consistent in the weighting used to derive all T_m* values. A 0 value is given for any residues at the terminal which do not have a full complement of 2 overlapping chains (*i.e.* for G_1 and X_1 of Chain A and G_1 of Chain B, as illustrated in Figure A-1). The 0 is interpreted by THe BuScr to imply that the neighbouring helicity type should be used to build these residues into the appropriate polypeptide chains.

A predicted T_m value for the entire triple-helical region of a collagen molecule can be readily obtained by averaging the values given in Table A-1, using eq. A-2 where necessary, for each triplet along the molecule's length. The T_m* for each residue in the structure can then be compared to the average T_m of the structure as a whole. For the purposes of modelling type I collagen herein, the triple-helical parameters are chosen using:

Parameters for residue	
IR	T_m^* (boxcar) $\geq T_m$ (overall average)
AR	$T_m^* < T_m$ (overall average)

(A-3)

For future use, this can be easily tweaked as desired. At present, this seems the most logical manner of selecting the helical parameters.

Testing THe BuScr with PDB entry 1BKV

It may not seem entirely likely that a simple two-state IR/AR model will provide a reliable model of the collagen triple-helix. Furthermore, why use something like T_m^{ave} as a test statistic (eq. A-3) to determine whether a given T_m* should imply IR or AR? To test out both the two-state model and the use of T_m^{ave}, a variety of comparisons were carried out relative to PDB entry 1BKV (Kramer *et al* 1999). This is currently the only high-resolution crystal structure of a triple-helical peptide containing both AR and IR regions. 1BKV contains three chains of sequence:

$(P\text{-}O\text{-}G)_3$-I-T-G-A-R-G-L-A-G-$(P\text{-}O\text{-}G)_4$

and was solved at 2.0 Å. In Rainey and Goh (2002), we used residues 10-21 to determine statistics for the AR category, with the remaining 19 residues used in the IR pool of statistics. In comparison to THe BuScr, one would expect a real triple-helix to have some subtle transition between IR and AR regions. Therefore, the ability to predict the structure of 1BKV is an excellent test for the methods used by THe BuScr.

Prediction using the sequence:

GPO-GPO-GPO-GIT-GAR-GLA-GPO-GPO-GPO-GPO

was carried out. For chain A, the N-terminal GP need to be removed from the prediction correspond directly to the resolved portion of 1BKV; likewise, the N-terminal G from chains B and C also needs to be removed. Finally, chains B and C show a large deviation at the final O of the prediction for all T_m cutoffs employed; correspondingly, these residues show high B-factors in the PDB file, an indication of lower certainty in their atom coordinates. Therefore, the final predictions are quoted both with and without this Hyp residue included for chains B and C.

For comparison purposes, the all-heavy atom RMSD is also given for the cutoff of T_m^{ave} as predicted using THe BuScr for Pro in X, Hyp in Y and Ala in X/Y positions along with SCWRL for all remaining side-chains.

RMSD details (backbone - N,C$^{\alpha}$,C) of THe BuScr prediction vs. 1BKV

T_m cutoff	AR region – chain A	RMSD – without final O [B & C]	RMSD – with final O [B & C]
44	PO-GIT-GAR-GLA-GPO-G	0.702 Å	0.856 Å
43.5	PO-GIT-GAR-GLA-GPO	0.676 Å	0.840 Å
T_m^{ave} = 43.13	O-GIT-GAR-GLA-GPO	0.622 Å *0.903 Å*[a]	0.799 Å
42.3	O-GIT-GAR-GLA-GP	0.620 Å	0.797 Å
42	GIT-GAR-GLA-GP	0.621 Å	0.797 Å
41	GIT-GAR-GLA-G	0.632 Å	0.813 Å
40	IT-GAR-GLA-G	0.654 Å	0.832 Å

a) RMSD for all heavy atoms using THe BuScr to predict Pro in X, Hyp in Y, and Ala in X/Y and SCWRL 3.0 (Canutescu et al 2003) for remainder of side-chains.

References

Bachinger, H.P., Davis, J.M. (1991) Sequence specific thermal stability of the collagen triple helix, *Int J Biol Macromol*, **13**, 152-156.

Chan, V.C., Ramshaw, J.A., Kirkpatrick, A., Beck, K., Brodsky, B. (1997) Positional preferences of ionizable residues in Gly-X-Y triplets of the collagen triple-helix, *J Biol Chem*, **272**, 31441-31446.

Kramer, R.Z., Bella, J., Mayville, P., Brodsky, B., Berman, H.M. (1999) Sequence-dependent conformational variations of collagen triple-helical structure. *Nat Struct Biol*, **6**, 454–457.

Krivov, G.G., Shapovalov, M.V., Dubrack Jr., R.L. (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**, 778-795.

Persikov, A.V., Ramshaw, J.A., Brodsky, B. (2000a) Collagen model peptides: Sequence dependence of triple-helix stability, *Biopolymers*, **55**, 436-450.

Persikov, A.V., Ramshaw, J.A., Kirkpatrick, A., Brodsky, B. (2000b) Amino acid propensities for the collagen triple-helix, *Biochemistry*, **39**, 14960-14967.

Persikov, A.V., Ramshaw, J.A., Kirkpatrick, A., Brodsky, B. (2002) Peptide investigations of pairwise interactions in the collagen triple-helix, *J Mol Biol*, **316**, 385-394.

Rainey, J.K. (2003) Collagen structure and preferential assembly explored by parallel microscopy and bioinformatics. *Ph.D. Thesis*, University of Toronto, Toronto, Ontario, Canada.

Rainey, J.K., Goh, M.C. (2002) A statistically derived parameterization for the collagen triple-helix. *Protein Sci*, **11**, 2748-2754.

Rainey, J.K., Goh, M.C. (2004). Statistically based reduced representation of amino acid side chains. *J Chem Inf Comput Sci,* **40**, 817-830**.**