

# REF3KG2\_master\_w-ADD\_432seqs

PDB: 3KG2 - May 25, 2023

## 1 SCA 6.1 - REF3KG2\_master\_w-ADD\_432seqs

```
[ ]: import os
import time
import matplotlib.pyplot as plt
import math
import numpy as np
import copy
import scipy.cluster.hierarchy as sch
from scipy.stats import scoreatpercentile
from pysca import scaTools as sca
import colorsys
# import mpld3
import pickle as pickle
from optparse import OptionParser

if not os.path.exists('../output/'):
    os.makedirs('../output/')

%matplotlib inline
```

Read in the results of the above three scripts (scaProcessMSA, scaCore and scaSectorID), stored as three dictionaries in the database REF3KG2\_master\_w-ADD\_432seqs.db. To see what variables are stored in each dictionary, use:

```
>>> list(db)
```

```
[ ]: db = pickle.load(open('/home/ibarnun/Projects/Jake_project/pySCA/output/
↳REF3KG2_master_w-ADD_432seqs.db', 'rb'))
Dseq = db['sequence']
Dsca = db['sca']
Dsect = db['sector']
```

### 1.0.1 I. Statistical Structure of the Multiple Sequence Alignment (MSA)

Plot a histogram of all pairwise sequence identities (*left panel*) and a global view of the sequence similarity matrix (defined by  $S \equiv \frac{1}{L}XX^T$ ) (*right panel*). From the matrix at right, it is clear that the alignment is composed of several distinct sequence families.

```
[ ]: # List all elements above the diagonal (i<j):
listS = [Dzca['simMat'][i,j] for i in range(Dzca['simMat'].shape[0]) \
        for j in range(i+1, Dzca['simMat'].shape[1])]

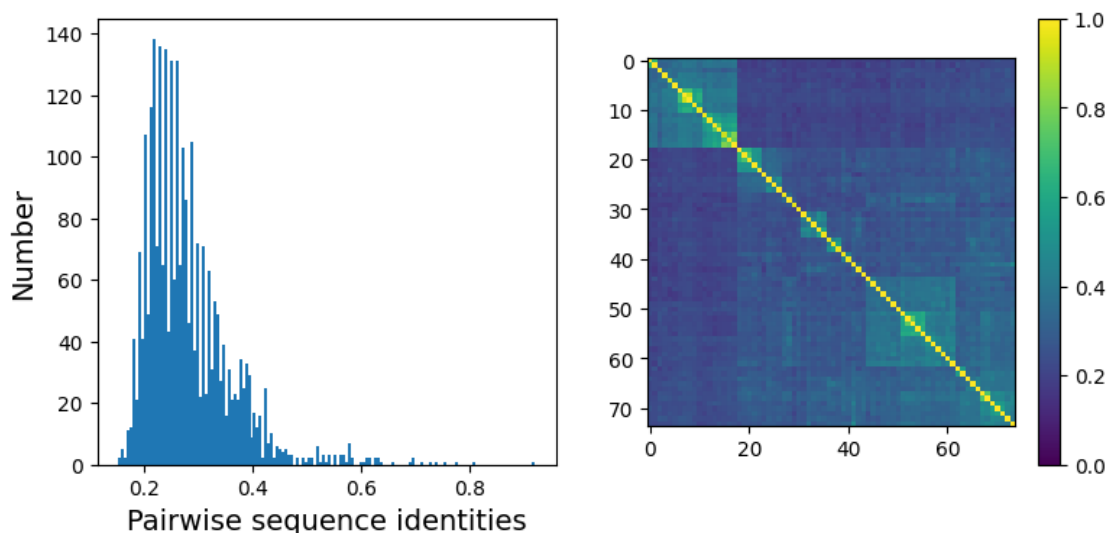
# Cluster the sequence similarity matrix
Z = sch.linkage(Dzca['simMat'], method = 'complete', metric = 'cityblock')
R = sch.dendrogram(Z,no_plot = True)
ind = R['leaves']
```

```
[ ]: # Plotting
plt.rcParams['figure.figsize'] = 9, 4

plt.subplot(121)
plt.hist(listS, math.floor(Dseq['Npos']/2))
plt.xlabel('Pairwise sequence identities', fontsize=14)
plt.ylabel('Number', fontsize=14)

plt.subplot(122)
plt.imshow(Dzca['simMat'][np.ix_(ind,ind)], vmin=0, vmax=1)
plt.colorbar()
```

```
[ ]: <matplotlib.colorbar.Colorbar at 0x7f4bae8ae7a0>
```



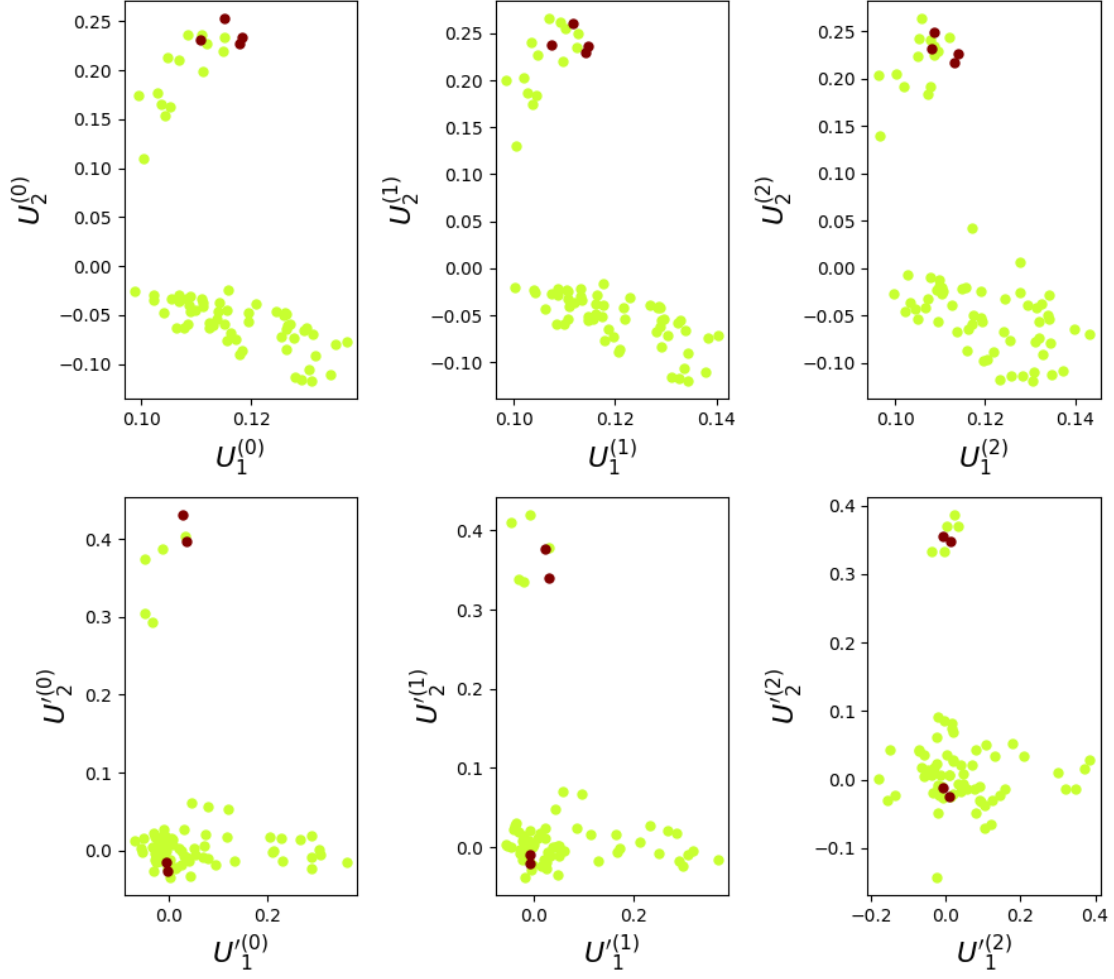
To examine the role of sequence and position weighting on the structure of the sequence space, we compute correlation matrices between all pairs of sequences, either with or without sequence and position weights and project the corresponding sequence space (by eigenvalue decomposition) down to a small set of top modes that contain the statistically dominant relationships between sequences. Since eigenvalue decomposition does not necessarily provide the best representation of sequence groups (for reasons described in “xx”), we also apply independent components analysis (or ICA)

to the top few eigenmodes; this manipulation provides a representation in which the top groupings of sequences in the alignment (if such exists) should separate along the so-called independent components (or ICs). Below we plot the following eigenmodes (*top row*) and independent components (*bottom row*):

- $U^{(0)}$  and  $U'^{(0)}$ , the top eigenmodes and ICs without any weights;
- $U^{(1)}$  and  $U'^{(1)}$  the top eigenmodes and ICs with sequences weights;
- $U^{(2)}$  and  $U'^{(2)}$  the top eigenmodes and ICs with both sequences and positional weights.

The sequences are colored by weight, with red indicating the most strongly downweighted sequences.

```
[ ]: Useq = Dsca['Useq']
      Uica = Dsca['Uica']
      plt.rcParams['figure.figsize'] = 9, 8
      ica = ["", "", "", "", "", "", ""]
      for k, U in enumerate(Useq+Uica):
          plt.subplot(2,3,k+1)
          sca.figWeights(U[:,0], U[:,1], Dseq['seqw'][0])
          plt.xlabel(r'${U}s}^{\{i\}}_1$(ica[k],k%3), fontsize=16)
          plt.ylabel(r'${U}s}^{\{i\}}_2$(ica[k],k%3), fontsize=16)
      plt.tight_layout()
```

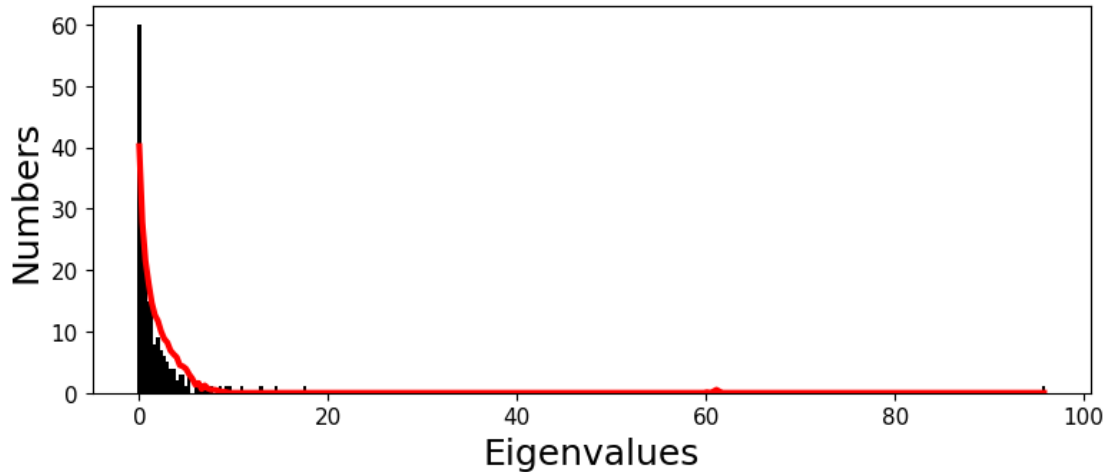


### 1.0.2 II. SCA conservation and coevolution

Plot the eigenspectrum of the SCA positional coevolution matrix ( $\tilde{C}_{ij}$ ) (*black bars*) and 10 trials of matrix randomization for comparison (*red line*). This graph is used to choose the number of significant eigenmodes.

```
[ ]: plt.rcParams['figure.figsize'] = 9, 3.5
hist0, bins = np.histogram(Dsca['Lrand'].flatten(), bins=Dseq['Npos'], \
                           range=(0,Dsect['Lsca'].max()))
hist1, bins = np.histogram(Dsect['Lsca'], bins=Dseq['Npos'], \
                           range=(0,Dsect['Lsca'].max()))
plt.bar(bins[:-1], hist1, np.diff(bins),color='k')
plt.plot(bins[:-1], hist0/Dsca['Ntrials'], 'r', linewidth=3)
plt.tick_params(labelsize=11)
plt.xlabel('Eigenvalues', fontsize=18); plt.ylabel('Numbers', fontsize=18);
print('Number of eigenmodes to keep is %i' %(Dsect['kpos']))
```

Number of eigenmodes to keep is 5



The data indicate generally good fits for the top 5 ICs, and we return the positions contributing to each IC in a format suitable for cut and paste into PyMol.

```
[ ]: plt.rcParams['figure.figsize'] = 10,5

Vpica = Dsect['Vpica']
for k in range(Dsect['kpos']):
    iqr = scoreatpercentile(Vpica[:,k],75) - scoreatpercentile(Vpica[:,k],25)
    binwidth=2*iqr*(len(Vpica)**(-0.33))
    nbins=int(round((max(Vpica[:,k])-min(Vpica[:,k]))/binwidth))
    plt.subplot(1,Dsect['kpos'],k+1)
    h_params = plt.hist(Vpica[:,k], nbins)
    x_dist = np.linspace(min(h_params[1]), max(h_params[1]), num=100)
    plt.plot(x_dist,Dsect['scaled_pd'][k], 'r',linewidth = 2)
    plt.xlabel(r'$V^p_{%i}$'%(k+1), fontsize=14)
    plt.ylabel('Number', fontsize=14)

for n,ipos in enumerate(Dsect['ics']):
    sort_ipos = sorted(ipos.items)
    ats_ipos = ([Dseq['ats'][s] for s in sort_ipos])
    ic_pymol = ('+'.join(ats_ipos))
    print('IC %i is composed of %i positions:' % (n+1,len(ats_ipos)))
    print(ic_pymol + "\n")
```

IC 1 is composed of 24 positions:

429+440+460+496+526+541+572+576+582+583+586+605+606+616+619+622+659+689+693+702+  
703+718+747+767

IC 2 is composed of 18 positions:

407+427+428+433+463+477+479+480+482+486+495+578+593+602+607+613+618+734

IC 3 is composed of 24 positions:

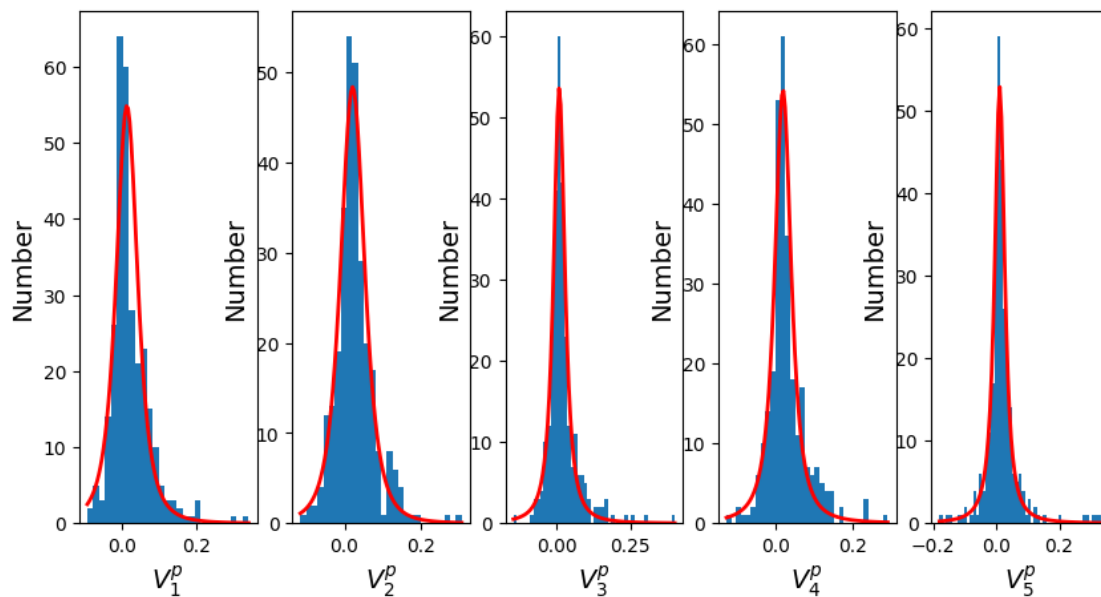
403+421+424+452+481+501+503+522+584+585+595+650+704+723+724+730+733+749+753+755+757+759+765+766

IC 4 is composed of 26 positions:

398+450+478+490+493+499+502+504+579+603+617+621+629+635+636+638+639+647+649+654+658+701+727+740+743+744

IC 5 is composed of 16 positions:

404+405+423+467+485+491+494+599+625+652+653+705+707+709+711+732



Sector 1: ICs 1, 2 and 4

Sector 2: ICs 3 and 5

```
[ ]: # plot the SCA positional correlation matrix, ordered by contribution to the
      ↪ top ICs
plt.rcParams['figure.figsize'] = 10, 10
plt.subplot(121)
plt.imshow(Dsca['Csca'][np.ix_(Dsect['sortedpos'], Dsect['sortedpos'])], \
           vmin=0, vmax=2, interpolation='none', aspect='equal', \
           extent=[0, sum(Dsect['icsize']), 0, sum(Dsect['icsize'])])
line_index=0
for i in range(Dsect['kpos']):
    plt.plot([line_index+Dsect['icsize'][i], line_index+Dsect['icsize'][i]], \
```

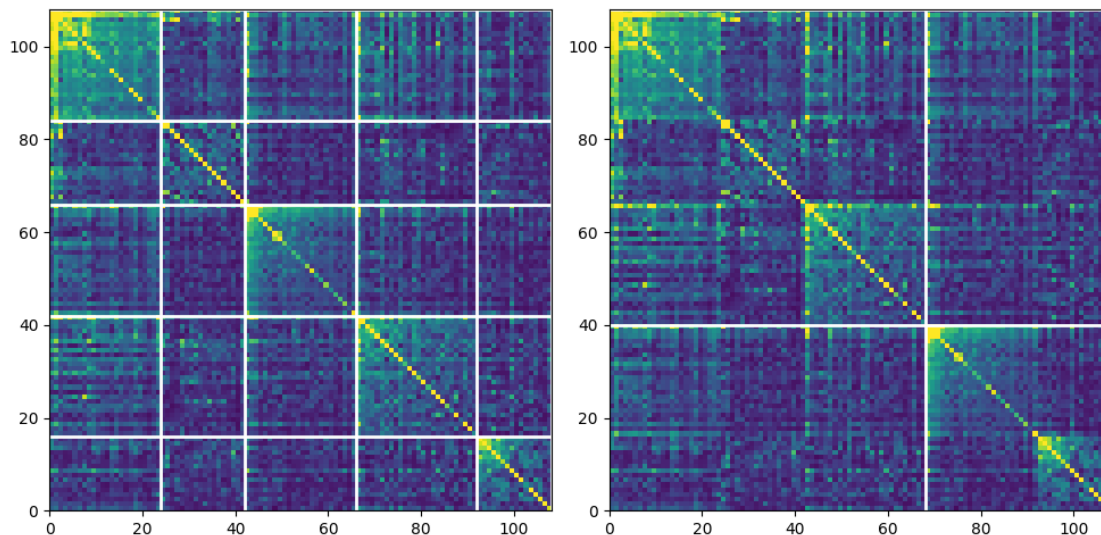
```

        [0,sum(Dsect['icsize'])], 'w', linewidth = 2)
plt.plot([0,sum(Dsect['icsize'])],[sum(Dsect['icsize'])-line_index,\
        sum(Dsect['icsize'])-line_index], 'w', linewidth = 2)
line_index += Dsect['icsize'][i]

#define the new sector groupings - 2 total
sec_groups = ([0,1,3],[2,4])
sectors = list()
for n,k in enumerate(sec_groups):
    s = sca.Unit()
    all_items = list()
    for i in k: all_items = all_items+Dsect['ics'][i].items
    s.items = all_items
    s.col = (1/len(sec_groups))*n
    sectors.append(s)

# plot the re-ordered matrix
plt.subplot(122)
line_index=0
sortpos = list()
for s in sectors:
    sortpos.extend(s.items)
plt.imshow(Dsca['Csca'][np.ix_(sortpos, sortpos)], vmin=0, vmax=2,\
        interpolation='none', aspect='equal',\
        extent=[0,len(sortpos),0,len(sortpos)])
for s in sectors:
    plt.plot([line_index+len(s.items),line_index+len(s.items)],\
        [0,len(sortpos)], 'w', linewidth = 2)
    plt.plot([0,sum(Dsect['icsize'])],[len(sortpos)-line_index, \
        len(sortpos)-line_index], 'w', linewidth=
↵= 2)
    line_index += len(s.items)
plt.tight_layout()

```



Print the sector positions, in a format suitable for pyMol, and create a pyMol session with the sectors (and decomposition into independent components) as separate objects.

```
[ ]: for i,k in enumerate(sectors):
    sort_ipos = sorted(k.items)
    ats_ipos = ([Dseq['ats'][s] for s in sort_ipos])
    ic_pymol = ('+'.join(ats_ipos))
    print('Sector %i is composed of %i positions:' % (i+1,len(ats_ipos)))
    print(ic_pymol + "\n")
sca.writePymol('../data/3kg2.pdb', sectors, Dsect['ics'], Dseq['ats'], \
               '../output/REF3KG2_master_w-ADD_432seqs.pml', 'A', '../data/', 0)
```

Sector 1 is composed of 68 positions:

398+407+427+428+429+433+440+450+460+463+477+478+479+480+482+486+490+493+495+496+  
499+502+504+526+541+572+576+578+579+582+583+586+593+602+603+605+606+607+613+616+  
617+618+619+621+622+629+635+636+638+639+647+649+654+658+659+689+693+701+702+703+  
718+727+734+740+743+744+747+767

Sector 2 is composed of 40 positions:

403+404+405+421+423+424+452+467+481+485+491+494+501+503+522+584+585+595+599+625+  
650+652+653+704+705+707+709+711+723+724+730+732+733+749+753+755+757+759+765+766