Merck Serono
Living science, transforming lives

MERCK

Development of web-based tools with RDKit & Django

Paul Czodrowski, Merck KGaA Darmstadt
Small Molecule Platform, Computational Chemistry
paul.czodrowski@merckgroup.com

RDKit User Group Meeting, London / October 2012

a 3-class aqueous solubility model

- Huuskonen dataset (http://dx.doi.org/10.1021/ci9901338 & http://www.cheminformatics.org/datasets/huuskonen/index.html)
- Training set: 1025 compounds
- Test set: 257 compounds
- Classification based on aqueous solubility

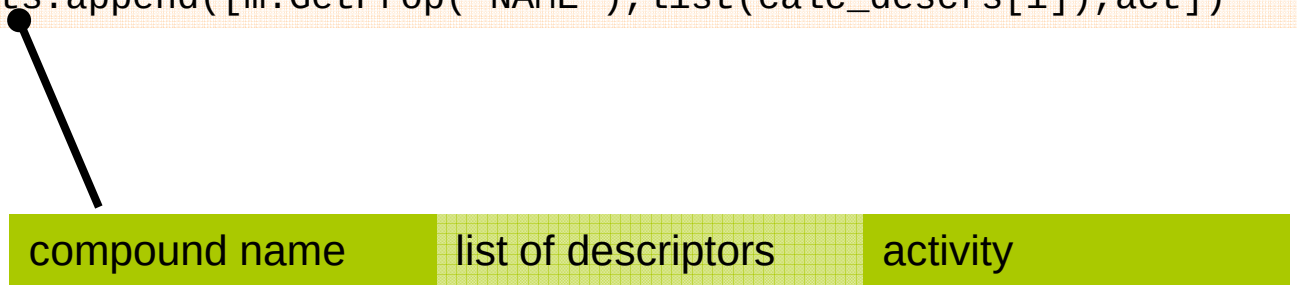
aqueous solubility	solubility class	# compounds
≤ -3.0	0 [low]	417
> -3.0 & ≤ -1.0	1 [medium]	402
> -1.0	2 [high]	206

code snippet – calculate the descriptors

fingerprint calculation

```
descr_list   = [x[0] for x in Descriptors._descList]
descr_list.remove('MolecularFormula')
descr_obj=MoleculeDescriptors.MolecularDescriptorCalculator(descr_list)
calc_descrs = [descriptor_object.CalcDescriptors(x) for x in cpds]

pts = []
for i,m in enumerate(ms):
    if m.GetProp('SOL_classification')=='(A) low':
        act=2
    elif m.GetProp('SOL_classification')=='(B) medium':
        act=1
    else:
        act=0
    pts.append([m.GetProp('NAME'),list(calc_descrs[i]),act])
```



code snippet – train the model

```
ndescrs = len(cpds_w_descr[0])-2
attrs = range(1,ndescrs+1)
nPossible = [0]+[2]*ndescrs+[2]
boundsPerVar = [0]+[1]*ndescrs+[0]
cmp = Composite()
cmp.Grow(pts,attrs=[1],nPossibleVals=[3],nTries=10,
         buildDriver=CrossValidate.CrossValidationDriver,
         treeBuilder=SigTreeBuilder,needsQuantization=False,maxDepth=3)
```

attrs	list of descriptors to use
nPossible	number of possible values of each descriptor
boundsPerVar	number of quantization bounds to automatically find for each descriptor

```
attrs        [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37,
 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75,
 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110,
111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 14
1, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171,
 172, 173, 174, 175]   175
nPossible    [0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2
, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]   177
boundsPerVar [0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0]   177
```

code snippet – model statistics

```
res = ScreenComposite.ShowVoteResults(range(len(pts)), pts, cmp, \
      3, 0,errorEstimate=True)
```

```
        *** Vote Results ***
misclassified: 245/1025 (%23.90)        245/1025 (%23.90)

average correct confidence:    0.9543
average incorrect confidence:  0.8496

        Results Table:

         353      57       1       |  84.45
          64     333     111       |  82.63
           0      12      94       |  45.41
      ------- ------- -------
       84.65   82.84   45.63
```

performance on the test set

experiment

prediction	class0	class1	class2
class0	89	17	0
class1	13	90	7
class2	0	8	33

overall accuracy: 83 %

code snippet – similarity search

fingerprints of training set

```
fps = [AllChem.GetMorganFingerprintAsBitVect(x,2,2048) for x in ms]
pts = []
for i,m in enumerate(ms):
    pts.append([m,m.GetProp('SOL_classification'),fps[i]])
    cPickle.dump(pts,file('cpds.pkl','wb+'))
```

similarity search

```
sol_trainset = cPickle.load('cpds.pkl", "rb"))

top3 = TopNContainer(3)
refFP = AllChem.GetMorganFingerprintAsBitVect(mol,2,2048)
hits = top3.GetPts()

for datasetmol in sol_trainset:
    sol_TL = datasetmol[1]
    fp = datasetmol[2]
    tani = DataStructs.TanimotoSimilarity(refFP,fp)
    top3.Insert((tani,sol_TL,datasetmol[0]))

hits = top3.GetPts()
hits.reverse()
```

how-to-set-up a simple website

```
urls.py
settings.py
manage.py
media/
templates/start_site.html
templates/results.html
webinter/views.py
```

development environment	productive environment
Comes with django installation	Apache/WSGI

pitfalls Django/WSGI/SLES

- our setup:
  - Apache server 2.2.12 on SLES 11SP1 (64 bit) with a separate virtualHost configured (to have it answering on and with an application-specific hostname)
  - python (compiled by user)
- `WSGIPythonHome` in `django.wsgi` not properly recognized
- `LD_LIBRARY_PATH` setting of original Apache shall not be modified
- necessary to run *distinct* Apache (`httpd-django.conf` besides `httpd.conf`)

Any experience with virtualenv?

django code snippets

urls.py

```
urlpatterns = patterns('webinter.views',
                       (r'^$', "jmecall"),
                       (r'^chime4sol/$', "chime4sol"),
                       (r'^get_sol/$', "get_sol"),
                       (r'^jme/$', "jmecall"),
                       (r'^draw_molecule/$', "draw_molecule"),
)
```

settings.py

```
FILE_CMP          = cPickle.load(file(os.path.join(MEDIA_ROOT,\
                                   'cmp.sol_model_descr.pkl'), "rb"))
FILE_CPDS         = cPickle.load(file(os.path.join(MEDIA_ROOT,\
                                   'cpds.MorganFP.pkl'), "rb"))
```

views.py

```
def get_sol(request):
    cmp = settings.FILE_CMP
    sol_trainset = settings.FILE_CPDS
    top3 = TopNContainer(3)
    errors = []
    if not 'q' in request.GET:
        errors.append('No query provided')
    else:
        query=request.GET['q']
    qj = QueryJob(inputtext=query,username=request.user.username)
    qj.setJobid()
    for word in query.split():
        mol = Chem.MolFromSmiles(str(query))
        mol.SetProp('_Name','nothing')
        #
        # nearest neighbor
        refFP = AllChem.GetMorganFingerprintAsBitVect(mol,2,2048)
        for datasetmol in sol_trainset:
            sol_TL = datasetmol[1]
            fp = datasetmol[2]
            tani = DataStructs.TanimotoSimilarity(refFP,fp)
            top3.Insert((tani,sol_TL,datasetmol[0]))
        hits = top3.GetPts()
        hit_list_dict = {}
        for cnt, sim_cpd in enumerate( hits ):
            skey = "image_query_cpd_%s" % cnt
            request.session[skey] = sim_cpd[2]
            if sim_cpd[1] == "(A) low":
                sol_TL = "0"
            elif sim_cpd[1] == "(B) medium":
                sol_TL = "1"
            else:
                sol_TL = "2"
```

results.html

```
{% for key, value in hit_list_dict.items %}
<tr  align=center>
    <td> <img src="/draw_molecule/?skey={{ value.3 }}" alt="huhu"></td>
    <td {% if value.2 == 0 %} bgcolor="orangered"
        {% else %} bgcolor="lawngreen" {% endif %}> {{ value.0}}</td>
    <td>{{ value.1 }}</td>
</tr>
{% endfor %}
```

Acknowledgments

- Daniel Kuhn
- Michael Krug
- Christian Griebel, Frank Morawietz, Trung Phan An
- Friedrich Rippmann
- Klaus Urbahns

- Greg Landrum
- RDKit mailing list!