An introduction to CHNOSZ

Jeffrey M. Dick

August 22, 2011

1 About

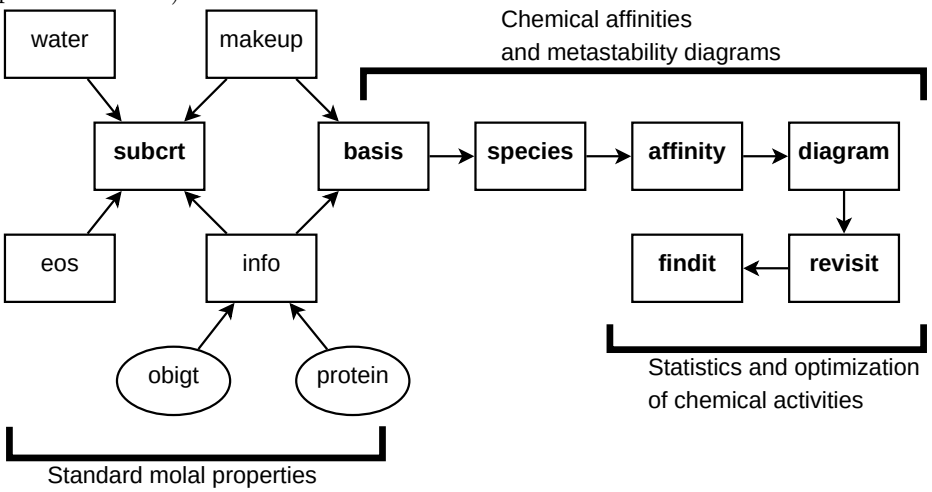
This document will orient you to the basic functionality of CHNOSZ, a package for the R software environment. R is a powerful language and also very fun to use. Don't worry if you're new to it; just plow through the examples below and you'll start to get the hang of it. If you want a more structured approach to learning the language, there are some excellent guides in the Manuals section of the R Project page. There is also a publication on CHNOSZ available [1], and a website.

The package was developed since 2006 to support a research project on the thermodynamic properties of proteins. Since that time, the functions in the package have expanded to include calculation of the thermodynamic properties of reactions, and especially the construction of equilibrium chemical activity diagrams for both inorganic and organic systems. The development of the package since 2009 has focused on the calculation of the equilibrium chemical activities of large numbers of proteins with applications in interpretation of metagenomic data and protein abundances in a variety of settings.

The database and functions are flexible in their use, allowing one to model the relative stabilities of proteins, minerals or aqueous species using very similar commands. Examples below are intended to demonstrate basic usage to new users.

2 Outline of workflow

CHNOSZ is made up of a set of functions and supporting datasets. The major components of the package are shown in the figure below, which is a modified version of the flowchart shown in Ref. [1] (boxes – functions; ellipses – datasets).



Some common usage scenarios are:

- using `info()` to search for species in the thermodynamic database
- using `subcrt()` to calculate the thermodynamic properties of species and reactions

- using the sequence `basis()`, `species()`, `affinity()`, `diagram()` to assign the basis species that define the dimensions of chemical composition in a system, define the species of interest for relative stability calculations, calculate the affinities of formation reactions of the species of interest under reference (non-equilibrium) conditions, and to transform the non-equilibrium affinities to equilibrium chemical activities and plot the results.
- using `revisit()` to calculate/plot statistics of the chemical activities of the species of interest and `findit()` to search for combinations of activities of basis species, temperature and/or pressure that optimize those statistics. (These features, first appearing in version 0.9-3 of the package, are not covered in this document.)

The functions are designed with an interactive setting in mind; you can use CHNOSZ without having to write your own scripts. The examples in this vignette are meant to portray a simple interactive session. However, as you become more familiar with CHNOSZ and R, you will probably find it helpful to save sequences of function calls that produce interesting results. The results can then be reproduced on demand by yourself or others with whom you might share your scripts.

3 Installing and loading CHNOSZ

If you have just installed R, and you are online, installing the CHNOSZ package should be as simple as selecting "Install packages from CRAN" or similar menu item in the R GUI or using the following command to start the package installation process. (If you are not online, you instead have to tell R to install the package from a local package file.)

```
> install.packages("CHNOSZ")
```

Then load the CHNOSZ package to make its functions and data available in your working session.

```
> library(CHNOSZ)
```

The rest of this document assumes that the CHNOSZ package is loaded.

4 Thermodynamic database

4.1 `info()` part I

So you want to know what are the standard molal thermodynamic properties and equations of state parameters of aqueous ethylene? Look no further than the `info()` function, which provides a convenient interface to retrieve entries from the thermodynamic database packaged with CHNOSZ.

```
> info("ethylene")
```

```
info: ethylene (C2H4) available in aq, gas.
info: 88 refers to ethylene, C2H4 aq (SH90, 4.Sep.87)
```

There are two species named "ethylene" in the database. Normally, `info()` gives preference to aqueous species if they exist, so in this case we find that aqueous ethylene is species number 88 in the database. Let's display this entry, now by giving the species index to the function.

```
> info(88)
```

```
       name abbrv formula state ref1 ref2     date     G    H    S   Cp    V     a1
88 ethylene  <NA>    C2H4    aq SH90 <NA> 4.Sep.87 19450 8570 28.7 62.5 45.5 0.7856
        a2      a3     a4   c1    c2  omega Z
88 1263.91 -1.8737 -33014 39.1 97000 -40000 0
```

If you were instead interested in the properties of the gas, you could run:

```
> info("ethylene", "gas")

info: 2698 refers to ethylene, C2H4 gas (Sho93, 22.Sep.93)
```

`info()` itself is used by other functions in the package. It prints output to the screen, but also returns a numeric value if it finds a species matching the search term. So, we can retrieve the properties of aqueous acetic acid without having to type in the species ID number.

```
> aadata <- info(info("acetic acid"))

info: acetic acid (C2H4O2) available in aq, liq.
info: 515 refers to acetic acid, C2H4O2 aq (Sho95, 6.Mar.92)

> print(aadata)

           name abbrv formula state  ref1 ref2     date      G       H    S    Cp     V
515 acetic acid  <NA>  C2H4O2    aq Sho95 <NA> 6.Mar.92 -94760 -116100 42.7 40.56 52.01
         a1    a2     a3     a4     c1     c2  omega Z
515 1.16198 521.8 2.5088 -29946 42.076 -15417 -15000 0
```

4.2 thermo$refs

The thermodynamic data and other parameters used by the functions, as well as system definitions provided by the user in an interactive session, are stored in a list object called `thermo`.

```
> summary(thermo)

        Length Class      Mode
opt     14     -none-     list
element  6     data.frame list
obigt   20     data.frame list
refs     5     data.frame list
buffers  4     data.frame list
protein 25     data.frame list
stress  65     data.frame list
groups  22     data.frame list
basis    0     -none-     NULL
species  0     -none-     NULL
water    0     -none-     NULL
water2   0     -none-     NULL
```

Within this list, the thermodynamic database is contained in a data frame (an R object that is like a matrix with named columns), `thermo$obigt`, and the references to the original sources of thermodynamic data in the literature are listed in `thermo$refs`. Many of the authors who are responsible for these data would be grateful if you cite them whenever these data are used in publications! Use the `browse.refs()` function without any arguments to show citation information for all of the references in a browser window. You can include a species index number to open the URL(s) associated with that entry in the database (this requires an Internet connection).

```
> browse.refs(88)

   browse.refs:  opening URL for SH90 (E. L. Shock and H. C. Helgeson, 1990)
```

4.3 `info()` part II

Want to know what acids are in the database?

```
> info("acid")

info: no match for acid.
info: similar species names, abbreviations, or formulas are:
 [1] "a-aminobutyric acid"     "formic acid"             "acetic acid"
 [4] "propanoic acid"          "n-butanoic acid"         "n-pentanoic acid"
 [7] "n-hexanoic acid"         "n-heptanoic acid"        "n-octanoic acid"
[10] "n-nonanoic acid"         "n-decanoic acid"         "n-undecanoic acid"
[13] "n-dodecanoic acid"       "n-benzoic acid"          "o-toluic acid"
[16] "m-toluic acid"           "p-toluic acid"           "oxalic acid"
[19] "malonic acid"            "succinic acid"           "glutaric acid"
[22] "adipic acid"             "pimelic acid"            "suberic acid"
[25] "azelaic acid"            "sebacic acid"            "glycolic acid"
[28] "lactic acid"             "2-hydroxybutanoic acid"  "2-hydroxypentanoic acid"
[31] "2-hydroxyhexanoic acid"  "2-hydroxyheptanoic acid" "2-hydroxyoctanoic acid"
[34] "2-hydroxynonanoic acid"  "2-hydroxydecanoic acid"  "aspartic acid"
[37] "glutamic acid"           "uracil"                  "citric acid"
[40] "2-methylpropanoic acid"  "metacinnabar"            "sanidine,high"
[43] "phosphoric acid"         "acetamide"               "nicotinamide,red"
[46] "nicotinamide,ox"         "n-tridecanoic acid"      "n-tetradecanoic acid"
[49] "n-pentadecanoic acid"    "n-hexadecanoic acid"     "n-heptadecanoic acid"
[52] "n-octadecanoic acid"     "n-nonadecanoic acid"     "n-eicosanoic acid"
[55] "hydrofluoric acid"       "butanoic acid"           "NicotinamideRed"
[58] "NicotinamideOx"
```

Here, `info()` couldn't find an exact match to a name, so it performed a fuzzy search. That's why "uracil" and "metacinnabar" show up above. If you really just want species whose names include the term "acid", you can add a placeholder character to narrow the search. (Note: don't use an underscore ("_") here because that character is reserved for names of proteins. Any other character will do; here we use a space.)

```
> info(" acid")

info: no match for  acid.
info: similar species names, abbreviations, or formulas are:
 [1] "a-aminobutyric acid"     "formic acid"             "acetic acid"
 [4] "propanoic acid"          "n-butanoic acid"         "n-pentanoic acid"
 [7] "n-hexanoic acid"         "n-heptanoic acid"        "n-octanoic acid"
[10] "n-nonanoic acid"         "n-decanoic acid"         "n-undecanoic acid"
[13] "n-dodecanoic acid"       "n-benzoic acid"          "o-toluic acid"
[16] "m-toluic acid"           "p-toluic acid"           "oxalic acid"
[19] "malonic acid"            "succinic acid"           "glutaric acid"
[22] "adipic acid"             "pimelic acid"            "suberic acid"
[25] "azelaic acid"            "sebacic acid"            "glycolic acid"
[28] "lactic acid"             "2-hydroxybutanoic acid"  "2-hydroxypentanoic acid"
[31] "2-hydroxyhexanoic acid"  "2-hydroxyheptanoic acid" "2-hydroxyoctanoic acid"
[34] "2-hydroxynonanoic acid"  "2-hydroxydecanoic acid"  "aspartic acid"
[37] "glutamic acid"           "citric acid"             "2-methylpropanoic acid"
[40] "phosphoric acid"         "n-tridecanoic acid"      "n-tetradecanoic acid"
[43] "n-pentadecanoic acid"    "n-hexadecanoic acid"     "n-heptadecanoic acid"
[46] "n-octadecanoic acid"     "n-nonadecanoic acid"     "n-eicosanoic acid"
[49] "hydrofluoric acid"       "butanoic acid"
```

The names of species other than proteins use (almost) exclusively lowercase letters. `info()` can also be used to search the text of the chemical formulas as they are entered in the database; the symbols for the elements always start with a capital letter. The example below lists the formulas of aqueous species, then minerals, that contain the symbol commonly used to represent the hydroxide group.

```
> info("(OH)")

info: no match for (OH).
info: similar species names, abbreviations, or formulas are:
 [1] "B(OH)3"                        "U(OH)+3"
 [3] "Ti(OH)4"                       "Pd(OH)2"
 [5] "U(OH)+2"                       "Ru(OH)+"
 [7] "Ru(OH)+2"                      "Rh(OH)+"
 [9] "Rh(OH)+2"                      "Pd(OH)+"
[11] "Pt(OH)+"                       "KAl3(OH)6(SO4)2"
[13] "Mg4Al2(Al2Si2)O10(OH)8"        "Mg2Al(AlSi)O5(OH)4"
[15] "KFe3(AlSi3)O10(OH)2"           "Mg7Si8O22(OH)2"
[17] "Mg48Si34O85(OH)62"             "Mg2(OH)2(CO3)*3H2O"
[19] "Cu3(OH)2(CO3)2"                "AlO(OH)"
[21] "Mg(OH)2"                       "K(MgAl)Si4O10(OH)2"
[23] "FeAl2SiO5(OH)2"                "Mg3Si2O5(OH)4"
[25] "Mg5Al(AlSi3)O10(OH)8"          "Ca2Al3Si3O12(OH)"
[27] "Fe2Fe(FeSi)O5(OH)4"            "Fe5Al(AlSi3)O10(OH)8"
[29] "Al2Si2O5(OH)4"                 "Na(Ca2Mg5)(AlSi7)O22(OH2)"
[31] "Ca2FeAl2Si3O12(OH)"            "Na(Ca2Fe5)(AlSi7)O22(OH)2"
[33] "(Fe5Al2)(Al2Si6)O22(OH)2"      "Na(Ca2Fe4Al)(Al2Si6)O22(OH)2"
[35] "(Ca2Fe5)Si8O22(OH)2"           "Al(OH)3"
[37] "Na2(Mg3Al2)Si8O22(OH)2"        "Fe3Si2O5(OH)4"
[39] "Fe7Si8O22(OH)2"                "Na(Ca2Fe4Fe)(Al2Si6)O22(OH)2"
[41] "Mg5(OH)2(CO3)4*4H2O"           "CaAl2Si2O7(OH)2*H2O"
[43] "Na(Ca2Mg4Fe)(Al2Si6)O22(OH)2" "Na2(Mg3Fe2)Si8O22(OH)2"
[45] "Cu2(OH)2(CO3)"                 "CaAl2(Al2Si2)O10(OH)2"
[47] "Fe3Si4O10(OH)2"                "KAl2(AlSi3)O10(OH)2"
[49] "NaAl2(AlSi3)O10(OH)2"          "Na(Ca2Mg4Al)(Al2Si6)O22(OH)2"
[51] "KFe3(AlSi3)O10(OH)O-"          "KMg3(AlSi3)O10(OH)2"
[53] "Ca2Al2Si3O10(OH)2"             "Al2Si4O10(OH)2"
[55] "Na2(CaMg5)Si8O22(OH)2"         "Na2(Fe3Fe2)Si8O22(OH)2"
[57] "Mg4Si6O15(OH)2(H2O)2*4H2O"     "Fe2Al9Si4O23(OH)"
[59] "Mg3Si4O10(OH)2"                "(Ca2Mg5)Si8O22(OH)2"
[61] "C2H4(OH)2"                     "C3H5(OH)3"
```

5 Proteins

5.1 protein()

There are few things more fun than calculating the standard molal Gibbs energy of formation from the elements at 25 °C and 1 bar of a protein using group additivity. And there are few proteins whose thermodynamic properties are more well studied than lysozyme from the egg of the chicken.

```
> protein("LYSC_CHICK")

  protein organism    ref  abbrv chains Ala Cys Asp Glu Phe Gly His Ile Lys Leu Met Asn
6    LYSC    CHICK BBA+03 P00698      1  12   8   7   2   3  12   1   6   6   8   2  14
  Pro Gln Arg Ser Thr Val Trp Tyr
6   2   3  11  10   7   6   6   3
```

```
> protein(6)

protein: found LYSC_CHICK (C613H959N193O185S10, 129 residues)
        name abbrv             formula state   ref1 ref2 date        G         H       S
1 LYSC_CHICK    NA C613H959N193O185S10    aq BBA+03   NA   NA -4206050 -10369700 4175.86
        Cp        V      a1     a2     a3     a4     c1     c2 omega Z
1 6415.553 10420.89 2512.58 345.88 450.87 -409.5 7768.7 -701.5 -7.94 0
```

What happened there? Well, the first line extracted the row (rownumber 6) of `thermo$protein` that contains the amino acid composition of LYSC_CHICK. The second line used group additivity [2] to calculate the standard molal thermodynamic properties and equations of state parameters of the aqueous protein species.

5.2 info()

Most of the time you probably won't be using the `protein()` function. That's because `info()` recognizes the underscore character as being an essential part of the name of a protein. The names of proteins in CHNOSZ are mostly consistent with those used in Swiss-Prot/UniProtKB.

```
> si <- info("LYSC_CHICK")

protein: found LYSC_CHICK (C613H959N193O185S10, 129 residues)
info: 2926 refers to LYSC_CHICK, C613H959N193O185S10 aq (BBA+03)

> info(si)

           name abbrv             formula state   ref1 ref2 date        G         H
2926 LYSC_CHICK  <NA> C613H959N193O185S10    aq BBA+03 <NA> <NA> -4206050 -10369700
           S       Cp        V      a1    a2     a3       a4     c1       c2   omega Z
2926 4175.86 6415.553 10420.89 251.258 34588 450.87 -4095000 7768.7 -7015000 -794000 0
```

When CHNOSZ is first loaded, the thermodynamic properties and parameters of the proteins are not present in `thermo$obigt`. Therefore, the first call to `info()` just above had a side effect of adding the computed properties and parameters to `thermo$obigt`.

6 Reaction properties

6.1 A single species

A major feature of CHNOSZ is the ability to calculate standard molal properties of species and reactions as a function of temperature and pressure. The function used is called `subcrt()`, which takes its name (with modification) from the well known SUPCRT package [3]. `subcrt()`, like `info()`, has the name of a species (including proteins) as its first argument (it also works if you give it the numeric index of the species in the database). If no reaction coefficients are given, the function calculates the standard molal properties of the indicated species on a default temperature-pressure grid.

```
> subcrt("water")

subcrt: 1 species at 15 values of T and P (wet)
$species
   name formula state ispecies
1 water     H2O   liq        1

$out
$out$water
        T          P       rho     logK         G         H        S        V       Cp
```

```
1    0.01   1.000000 0.9998289 45.03529 -56289.50 -68767.75 15.13238 18.01828 18.20559
2   25.00   1.000000 0.9970614 41.55247 -56687.71 -68316.76 16.71228 18.06830 18.01160
3   50.00   1.000000 0.9880295 38.63281 -57123.89 -67866.54 18.16234 18.23346 18.00464
4   75.00   1.000000 0.9748643 36.15435 -57594.93 -67416.13 19.50485 18.47970 18.04163
5  100.00   1.013220 0.9583926 34.02698 -58098.40 -66963.78 20.75956 18.79731 18.15793
6  125.00   2.320144 0.9390726 32.18315 -58631.71 -66507.34 21.94192 19.18403 18.33334
7  150.00   4.757169 0.9170577 30.57178 -59193.26 -66045.55 23.06398 19.64456 18.56643
8  175.00   8.918049 0.8923427 29.15313 -59781.38 -65576.63 24.13602 20.18866 18.88296
9  200.00  15.536499 0.8647434 27.89596 -60394.50 -65097.99 25.16818 20.83300 19.32884
10 225.00  25.478603 0.8338733 26.77533 -61031.25 -64605.89 26.17117 21.60424 19.97039
11 250.00  39.736493 0.7990719 25.77115 -61690.35 -64095.00 27.15694 22.54515 20.91232
12 275.00  59.431251 0.7592362 24.86701 -62370.65 -63557.52 28.14000 23.72806 22.35126
13 300.00  85.837843 0.7124075 24.04945 -63071.13 -62980.94 29.14072 25.28777 24.73943
14 325.00 120.457572 0.6545772 23.30725 -63790.84 -62341.39 30.19520 27.52189 29.44748
15 350.00 165.211289 0.5746875 22.63103 -64528.89 -61575.58 31.39713 31.34782 43.59852
```

The columns in the output are temperature (°C), pressure (bar), density of water (g cm^{-3}), logarithm of the equilibrium constant (only meaningful for reactions; see below), and standard molal Gibbs energy and enthalpy of formation from the elements (cal mol^{-1}), and standard molal entropy (cal K^{-1} mol^{-1}), volume (cm^3 mol^{-1}) and heat capacity (cal K^{-1} mol^{-1}).

Compared to other species available in CHNOSZ, the equations for calculating the properties of liquid H_2O are quite complex. The package uses a Fortran subroutine taken from SUPCRT for these calculations. See `help(water)` for more information.

6.2 A reaction

To calculate the properties of a reaction, enter the stoichiometric reaction coefficients as a second argument to `subcrt()`. Reactants have negative coefficients, and products have positive coefficients. The function call below also shows the specification of temperature.

```
> subcrt(c("C2H5OH", "O2", "CO2", "H2O"), c(-1, -3, 2, 3), T = 37)

subcrt: 4 species at 310.15 K and 1 bar (wet)
$reaction
     coeff    name formula state ispecies
112     -1 ethanol  C2H5OH    aq      112
2691    -3  oxygen      O2   gas     2691
69       2     CO2     CO2    aq       69
1        3   water     H2O   liq        1

$out
   T P       rho     logK         G         H         S        V      Cp
1 37 1 0.9933251 218.6729 -310330.2 -333262.2 -73.89356 67.43932 67.1269
```

For historical reasons (i.e., the prevalence of the use of oxygen fugacity in geochemical modeling [4]), O_2 breaks the general rule in CHNOSZ that species whose states are not specified are given the aqueous designation if it is available in the thermodynamic database. If you want to specify the physical states of the species in the reaction, that's possible too. For example, we can ensure that dissolved O_2 instead of the gaseous form is used in the calculation.

```
> subcrt(c("C2H5OH", "O2", "CO2", "H2O"), c(-1, -3, 2, 3), c("aq", "aq",
+     "aq", "liq"), T = 37)

subcrt: 4 species at 310.15 K and 1 bar (wet)
$reaction
    coeff    name formula state ispecies
```

```
112    -1 ethanol  C2H5OH    aq      112
67     -3      O2      O2    aq       67
69      2     CO2     CO2    aq       69
1       3   water     H2O   liq        1

$out
   T P       rho     logK       G         H         S         V        Cp
1 37 1 0.9933251 227.5908 -322986 -326236.2 -10.43305 -26.46463 -66.17582
```

A useful feature of `subcrt()` is that it emits a warning if the reaction is not balanced. Let's say you forgot to account for oxygen on the left-hand side of the reaction[1].

```
> subcrt(c("C2H5OH", "CO2", "H2O"), c(-1, 2, 3), T = 37)

subcrt: 3 species at 310.15 K and 1 bar (wet)
subcrt: reaction is not balanced; it is missing this composition:
  O
 -6
$reaction
    coeff    name formula state ispecies
112    -1 ethanol  C2H5OH    aq      112
69      2     CO2     CO2    aq       69
1       3   water     H2O   liq        1

$out
   T P       rho     logK         G       H        S        V       Cp
1 37 1 0.9933251 219.9202 -312100.3 -333009 74.02581 67.43932 88.28986
```

The function still reports the results of the calculations, but use them very cautiously (only if you have a specific reason for writing an unbalanced reaction). In the next section we'll see how to use another feature of CHNOSZ to automatically balance reactions.

7 Basis species

7.1 What are basis species?

Basis species are a minimal number of chemical species that represent the compositional variation in a system. Operationally, a **system** is the combination of basis species and species of interest which is set up by the user to investigate a real-life system. The basis species are akin to thermodynamic components, but can include charged species.

There are at least two reasons to define the basis species when using CHNOSZ. First, you might want to use them to automatically balance reactions. Second, they are required for making chemical activity diagrams. Let's start with an example that *doesn't* work.

```
> basis(c("CO2", "H2O", "NH3", "H2S", "H+"))



    Error in put.basis(basis, mystates) :
      the stoichiometric matrix must be square and invertible
    In addition: Warning messages:
    1: basis: 5 compounds ( CO2 H2O NH3 H2S H+ )
    2: basis: 6 elements ( C H N O S Z )
```

[1]This example is motivated by the unbalanced reaction found at the Wikipedia entry on ethanol metabolism on 2010-09-23 and still present as of 2011-08-15: "Complete Reaction: C_2H_6O(Ethanol)→C_2H_4O(Acetaldehyde)→$C_2H_4O_2$(acetic Acid) →Acetyl-CoA→$3H_2O$+$2CO_2$".

A limitation of CHNOSZ is that the number of basis species must be equal to the number of elements, plus one if charge is present. This way, any possible species of interest made up of these elements can be compositionally represented by a linear combination of the basis species. Now let's write a working basis definition.

```
> basis(c("CO2", "H2O", "NH3", "O2", "H2S", "H+"))

    C H N O S Z ispecies logact state
CO2 1 0 0 2 0 0       69      0    aq
H2O 0 2 0 1 0 0        1      0   liq
NH3 0 3 1 0 0 0       68      0    aq
O2  0 0 0 2 0 0     2691      0   gas
H2S 0 2 0 0 1 0       70      0    aq
H+  0 1 0 0 0 1        3      0    aq
```

First basis definition! Note the column names, which give CHNOSZ its name. These represent the elements in the commonly-occurring amino acids, together with charge, denoted by "Z".

7.2 Auto-balancing a reaction

Now that the basis species are defined, try the unbalanced reaction again.

```
> subcrt(c("C2H5OH", "CO2", "H2O"), c(-1, 2, 3), T = 37)

subcrt: 3 species at 310.15 K and 1 bar (wet)
subcrt: reaction is not balanced; it is missing this composition:
  O
 -6
subcrt: adding missing composition from basis definition and restarting...
subcrt: 4 species at 310.15 K and 1 bar (wet)
$reaction
     coeff    name formula state ispecies
112     -1 ethanol  C2H5OH    aq      112
69       2     CO2     CO2    aq       69
1        3   water     H2O   liq        1
2691    -3  oxygen      O2   gas     2691

$out
   T P     logK         G         H         S        V      Cp
1 37 1 218.6729 -310330.2 -333262.2 -73.89356 67.43932 67.1269
```

Here, `subcrt()` detected an unbalanced reaction, but since the missing element was among the elements of the basis species, it added the appropriate amount of $O_{2(gas)}$ to the reaction before running the calculations. You can go even further and eliminate CO_2 and H_2O from the function call, but still get the same results.

```
> subcrt(c("C2H5OH"), c(-1), T = 37)

subcrt: 1 species at 310.15 K and 1 bar (wet)
subcrt: reaction is not balanced; it is missing this composition:
 C H O
 2 6 1
subcrt: adding missing composition from basis definition and restarting...
subcrt: 4 species at 310.15 K and 1 bar (wet)
$reaction
     coeff    name formula state ispecies
112     -1 ethanol  C2H5OH    aq      112
```

```
69       2     CO2     CO2    aq       69
1        3   water     H2O   liq        1
2691    -3  oxygen      O2   gas     2691

$out
   T P     logK         G         H         S        V      Cp
1 37 1 218.6729 -310330.2 -333262.2 -73.89356 67.43932 67.1269
```

What if you were interested in the thermodynamic properties of the reaction of ethanol to acetaldehyde, but didn't want to balance the reaction yourself (and you also didn't know how the formulas of the species are written in the database)?

```
> subcrt(c("ethanol", "acetaldehyde"), c(-1, 1), T = 37)

subcrt: 2 species at 310.15 K and 1 bar (wet)
subcrt: reaction is not balanced; it is missing this composition:
 H
 2
subcrt: adding missing composition from basis definition and restarting...
subcrt: 4 species at 310.15 K and 1 bar (wet)
$reaction
     coeff         name formula state ispecies
112   -1.0      ethanol  C2H5OH    aq      112
256    1.0 acetaldehyde  CH3CHO    aq      256
1      1.0        water     H2O   liq        1
2691  -0.5       oxygen      O2   gas     2691

$out
   T P     logK         G        H         S        V       Cp
1 37 1 32.90437 -46696.31 -50194.4 -11.18783 7.708236 -11.8625
```

Notice how 2 H's needed to be added to the right-hand side of the reaction; in our definition of basis species this comes out to $H_2O - 0.5O_2$. With a different choice of basis species, but the same elements, the reaction might look quite different. As an example, suppose you had amino acids in mind. The first line below, `data(thermo)`, is a quick way to reset the `thermo` object to its original state, in order to clear the current system definition.

```
> data(thermo)
> basis(c("glutamic acid", "methionine", "isoleucine", "lysine", "tyrosine",
+     "H+"))

          C  H N O S Z ispecies logact state
C5H9NO4   5  9 1 4 0 0     1514      0    aq
C5H11NO2S 5 11 1 2 1 0     1525      0    aq
C6H13NO2  6 13 1 2 0 0     1520      0    aq
C6H14N2O2 6 14 2 2 0 0     1522      0    aq
C9H11NO3  9 11 1 3 0 0     1531      0    aq
H+        0  1 0 0 0 1        3      0    aq

> subcrt(c("ethanol", "acetaldehyde"), c(-1, 1), T = 37)

subcrt: 2 species at 310.15 K and 1 bar (wet)
subcrt: reaction is not balanced; it is missing this composition:
 H
 2
subcrt: adding missing composition from basis definition and restarting...
```

```
subcrt: 5 species at 310.15 K and 1 bar (wet)
$reaction
      coeff         name   formula state ispecies
112  -1.000      ethanol    C2H5OH    aq      112
256   1.000 acetaldehyde    CH3CHO    aq      256
1520  0.500   isoleucine  C6H13NO2    aq     1520
1522 -0.125       lysine C6H14N2O2    aq     1522
1531 -0.250     tyrosine  C9H11NO3    aq     1531

$out
   T P      logK        G        H          S         V        Cp
1 37 1 -1.341659 1904.018 1703.277 -0.5291135 -2.397983 -10.05446
```

In this case, the function finds that 2 H's are the compositional equivalent of $0.5C_6H_{13}NO_2 - 0.125C_6H_{14}N_2O_2 - 0.250C_9H_{11}NO_3$. It's pretty easy for the computer to figure that out using matrix operations, but probably isn't something you'd want to do by hand. You might complain that this reaction is not likely to represent an actual metabolic process ... as always, the challenge (and fun) of coming up with a useful basis definition is in relating the species to observable quantities.

7.3 It works for proteins too!

Let's set the basis definition again, this time using a keyword that refers to a preset combination of basis species commonly encountered in the documentation for CHNOSZ. Then we will use `subcrt()` to calculate the thermodynamic properties of a reaction to form a protein from the basis species.

```
> data(thermo)
> basis("CHNOS+")

    C H N O S Z ispecies logact state
CO2 1 0 0 2 0 0       69     -3    aq
H2O 0 2 0 1 0 0        1      0   liq
NH3 0 3 1 0 0 0       68     -4    aq
H2S 0 2 0 0 1 0       70     -7    aq
O2  0 0 0 2 0 0     2691    -80   gas
H+  0 1 0 0 0 1        3     -7    aq

> subcrt("LYSC_CHICK", 1, T = 25)

protein: found LYSC_CHICK (C613H959N193O185S10, 129 residues)
subcrt: 1 species at 298.15 K and 1 bar (wet)
subcrt: reaction is not balanced; it is missing this composition:
    C    H    N    O   S
 -613 -959 -193 -185 -10
subcrt: adding missing composition from basis definition and restarting...
subcrt: 6 species at 298.15 K and 1 bar (wet)
$reaction
      coeff       name             formula state ispecies
2926    1.0 LYSC_CHICK C613H959N193O185S10    aq     2926
69   -613.0        CO2                 CO2    aq       69
1    -180.0      water                 H2O   liq        1
68   -193.0        NH3                 NH3    aq       68
70    -10.0        H2S                 H2S    aq       70
2691  610.5     oxygen                  O2   gas     2691

$out
   T P      logK        G        H        S         V        Cp
1 25 1 -46799.28 63845637 66394946 8600.944 -18320.13 -27314.51
```

Note that using the keyword argument in `basis()` also set the logarithms of activities (or fugacity in the case of $O_{2(g)}$) to nominal values. While these settings do not affect the results of the `subcrt()` calculation (which normally returns only the standard molal properties of the reaction), they are essential for calculating the relative stabilities of the species of interest.

If the protein is not found in CHNOSZ's own database, the amino acid composition of the protein can be retrieved from the UniProt Knowledge Base using the Swiss-Prot name (if the computer is connected to the Internet). This is the only time a function in CHNOSZ asks for confirmation from a user, in order to give fair warning that an online activity is about to be performed.

```
> subcrt("ALAT1_HUMAN", 1, T = 25)
```

```
Shall I try an online search for  ALAT1_HUMAN _ SWISS ? y
protein: trying http://www.uniprot.org/uniprot/ALAT1_HUMAN ... got it!
protein: found P24298 ...  Alanine aminotransferase 1  (length 496).
protein: found ALAT1_HUMAN (C2429H3866N684O705S22, 496 residues)
subcrt: 1 species at 298.15 K and 1 bar (wet)
subcrt: reaction is not balanced; it is missing this composition:
     C     H    N    O   S
 -2429 -3866 -684 -705 -22
subcrt: adding missing composition from basis definition and restarting...
subcrt: 6 species at 298.15 K and 1 bar (wet)
$reaction
     coeff        name               formula state ispecies
2926     1 ALAT1_HUMAN C2429H3866N684O705S22    aq     2926
69   -2429         CO2                   CO2    aq       69
1     -885       water                   H2O   liq        1
68    -684         NH3                   NH3    aq       68
70     -22         H2S                   H2S    aq       70
2691  2519      oxygen                    O2   gas     2691
$out
   T P      logK         G         H        S      V        Cp
1 25 1 -191972.3 261897066 273248830 38245.59 -73411 -107650.2
```

8 Activity diagrams

8.1 Quick example: stability diagram for proteins

Suppose that we are asked to calculate the relative stabilities of some proteins from different organisms. We will use part of a case study presented in Ref. [1]. *Methanocaldococcus jannaschii* is a hyperthermophilic methanogen known to live at higher temperatures than *Methanococcus voltae* (also a methanogen) and *Haloarcula japonica* (a halophile). These archaeal organisms produce cell-surface glycoproteins (a.k.a. surface-layer proteins).

After defining the basis species we can define the **species of interest**, i.e. those proteins whose relative stabilities we wish to calculate.

```
> species(c("CSG_METJA", "CSG_METVO", "CSG_HALJP"))
```

```
protein: found CSG_METJA (C2555H4032N640O865S14, 530 residues)
protein: found CSG_METVO (C2575H4097N645O884S11, 553 residues)
protein: found CSG_HALJP (C3669H5647N971O1488, 828 residues)
   CO2  H2O NH3 H2S      O2 H+ ispecies logact state      name
1 2555 1042 640  14 -2643.5  0     2927     -3    aq CSG_METJA
2 2575 1070 645  11 -2668.0  0     2928     -3    aq CSG_METVO
3 3669 1367 971   0 -3608.5  0     2929     -3    aq CSG_HALJP
```

Note the output: the matrix denotes the coefficients of each of the basis species in the formation reaction for one mole of each of the species of interest. The **formation reaction** is the chemical reaction to form one mole of a species of interest (as a product) from a combination of basis species (as reactants and/or products, depending on the stoichiometric constraints). The formation reactions generally are *not* statements about the mechanisms of reactions. The species definition also includes reference values for the chemical activities of the species of interest.

Now we are all set up to calculate the chemical affinities of the formation reactions. The chemical affinity is the negative of the Gibbs energy change of a reaction per unit of reaction progress; it is calculated in CHNOSZ using $\boldsymbol{A} = 2.303RT\log(K/Q)$ (R – gas constant, T – temperature, K – equilibrium constant, Q – activity product).

`affinity()` can accept arguments describing the range of chemical conditions we're interested in. The names of the arguments can refer to the basis species. Here, we vary the logarithm of the fugacity of oxygen. The chemical activities of the other basis species are taken to be constants equal to the values shown above.

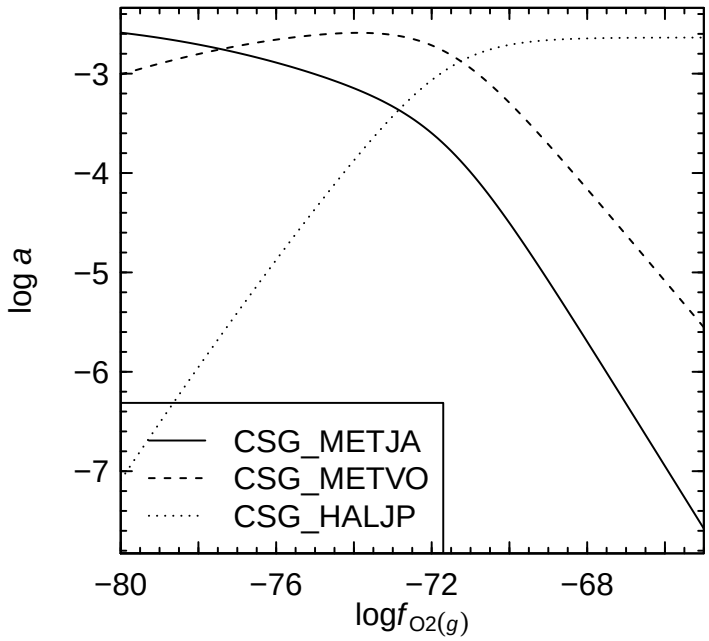
```
> a <- affinity(O2 = c(-80, -65))
```

```
affinity: temperature is 25 C
energy.args: pressure is Psat
energy.args: variable 1 is O2 at 128 increments from -80 to -65
affinity: loading ionizable protein groups
subcrt: 26 species at 298.15 K and 1 bar (wet)
```

Now we can use `diagram()` to plot the relative stabilities of the proteins in the system. We'll also specify where the legend should be placed on the plot.

```
> diagram(a, legend.x = "bottomleft")
```

```
diagram: immobile component is protein backbone group
diagram: conservation coefficients are 530 553 828
diagram: using residue equivalents
diagram: log total activity of PBB (from species) is 0.2812607
```



Notably, the protein from the organism found at the highest temperatures is relatively stable at more reduced conditions.

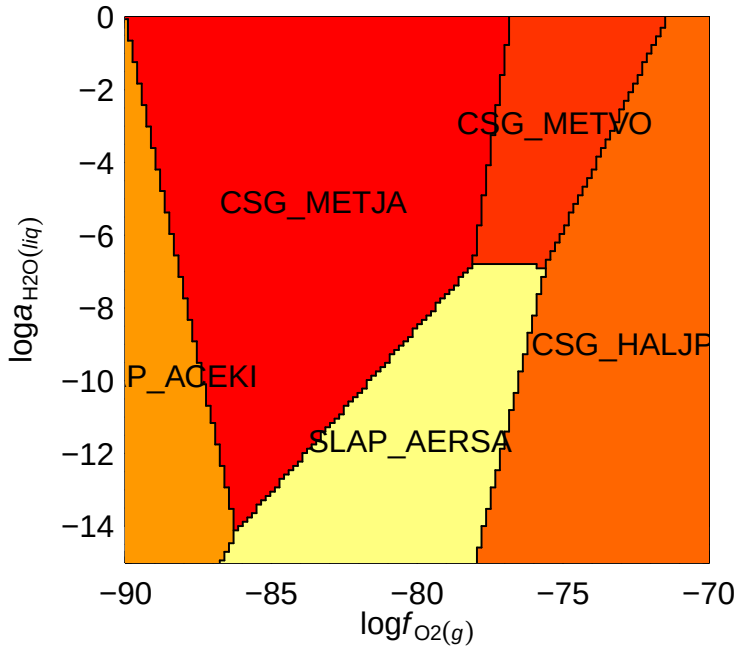
8.2 How does this work?

Here is a partial explanation: You use `affinity()` to calculate the chemical affinities of the formation reactions of the proteins, taking into account chemical activities of the proteins that are set to reference, non-equilibrium values. Then, the `diagram()` function transforms these non-equilibrium affinities into chemical activities of the proteins at metastable equilibrium (this is actually achieved using the Boltzmann distribution). These activities satisfy the conditions that 1) the total activity of an immobile component (for proteins, this defaults to the protein backbone group) is constant and 2) the chemical affinities of the formation reactions are all equal (but generally not zero). More details can be found in another vignette ("`protactiv`").

8.3 More proteins, more dimensions

Now let's add some bacterial surface-layer proteins. They are in some way functional analogs (but not homologs) of the archaeal cell-surface glycoproteins.

```
> species(c("SLAP_ACEKI", "SLAP_BACST", "SLAP_BACLI", "SLAP_AERSA"))

protein: found SLAP_ACEKI (C3584H5648N926O1138S4, 736 residues)
protein: found SLAP_BACST (C5676H9113N1489O1863S3, 1198 residues)
protein: found SLAP_BACLI (C3977H6396N1068O1286S2, 844 residues)
protein: found SLAP_AERSA (C2250H3580N618O716S2, 481 residues)
   CO2  H2O  NH3 H2S      O2 H+ ispecies logact state       name
1 2555 1042  640  14 -2643.5  0     2927     -3    aq  CSG_METJA
2 2575 1070  645  11 -2668.0  0     2928     -3    aq  CSG_METVO
3 3669 1367  971   0 -3608.5  0     2929     -3    aq  CSG_HALJP
4 3584 1431  926   4 -3730.5  0     2930     -3    aq SLAP_ACEKI
5 5676 2320 1489   3 -5904.5  0     2931     -3    aq SLAP_BACST
6 3977 1594 1068   2 -4131.0  0     2932     -3    aq SLAP_BACLI
7 2250  861  618   2 -2322.5  0     2933     -3    aq SLAP_AERSA

> basis(c("NH3", "H2S"), c(-1, -10))

    C H N O S Z ispecies logact state
CO2 1 0 0 2 0 0       69     -3    aq
H2O 0 2 0 1 0 0        1      0   liq
NH3 0 3 1 0 0 0       68     -1    aq
H2S 0 2 0 0 1 0       70    -10    aq
O2  0 0 0 2 0 0     2691    -80   gas
H+  0 1 0 0 0 1        3     -7    aq

> a <- affinity(O2 = c(-90, -70), H2O = c(-15, 0))

affinity: temperature is 25 C
energy.args: pressure is Psat
energy.args: variable 1 is O2 at 128 increments from -90 to -70
energy.args: variable 2 is H2O at 128 increments from -15 to 0
affinity: loading ionizable protein groups
subcrt: 30 species at 298.15 K and 1 bar (wet)

> diagram(a)

diagram: immobile component is protein backbone group
diagram: conservation coefficients are 530 553 828 736 1198 844 481
diagram: using residue equivalents
```



Equilibrium predominances for proteins as a function of two chemical activities! If you don't like the colors in the plot, don't worry... the colors can be changed by using the `col` argument of `diagram()`. This example hints at the multidimensional nature of the stability problem. Note how the order of predominance fields at $\log a_{H_2O} = 0$ matches the order of proteins with highest equilibrium activities in the previous diagram. Interpreting the meaning of low activities of H_2O in these calculations remains a challenge.

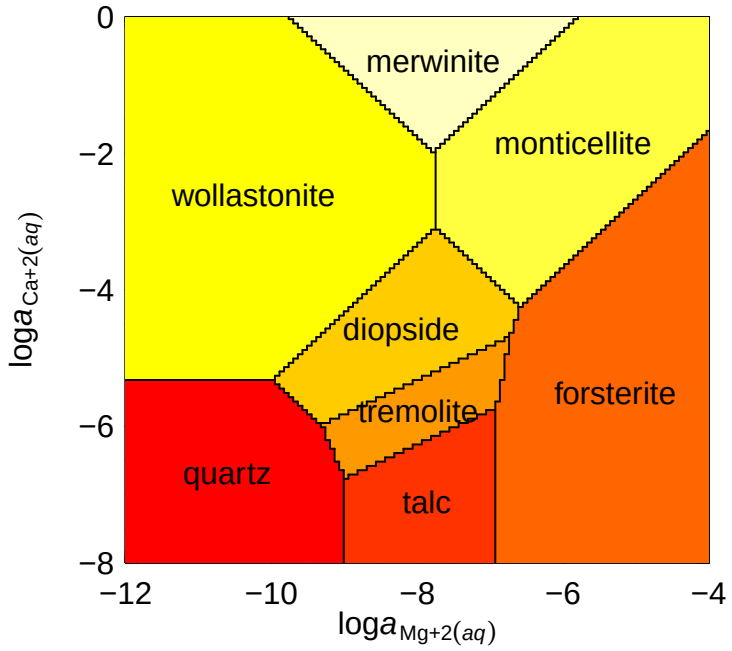
Why did we increase the activity of NH_3 and decrease that of H_2S? It was done here in order to increase the size of the equilibrium predominance fields of the bacterial proteins. This behavior is a result of the elemental makeup of the proteins: the bacterial proteins under consideration are, per residue, more nitrogen-rich and sulfur-poor than their archaeal counterparts (except for CSG_HALJP, which has no sulfur). CHNOSZ has a function to display the compositional makeup of the proteins, per residue, from the basis species.

```
> residue.info()

affinity: temperature is 25 C
energy.args: pressure is Psat
affinity: loading ionizable protein groups
subcrt: 30 species at 298.15 K and 1 bar (wet)
       CO2      H2O      NH3         H2S        O2          H+       name
1 4.820755 1.966038 1.207547 0.026415094 -4.987736 -0.10541556  CSG_METJA
2 4.656420 1.934901 1.166365 0.019891501 -4.824593 -0.10138353  CSG_METVO
3 4.431159 1.650966 1.172705 0.000000000 -4.358092 -0.22278182  CSG_HALJP
4 4.869565 1.944293 1.258152 0.005434783 -5.068614 -0.01766060 SLAP_ACEKI
5 4.737896 1.936561 1.242905 0.002504174 -4.928631 -0.01312341 SLAP_BACST
6 4.712085 1.888626 1.265403 0.002369668 -4.894550 -0.01438826 SLAP_BACLI
7 4.677755 1.790021 1.284823 0.004158004 -4.828482 -0.01992716 SLAP_AERSA
```

8.4 A mineral example

This example is modeled after a figure on p. 246 of Bowers et al., 1984 [5] for the system HCl-H2O-CaO-CO2-MgO-(SiO2) at 300 °C and 1000 bar.

```
> basis(c("HCl", "H2O", "Ca+2", "CO2", "Mg+2", "SiO2", "O2", "H+"), c(999,
+     0, 999, 999, 999, 999, 999, -7))

     C Ca Cl H Mg O Si Z ispecies logact state
HCl  0  0  1 1  0 0  0 0      883    999    aq
H2O  0  0  0 2  0 1  0 0        1      0   liq
Ca+2 0  1  0 0  0 0  0 2       10    999    aq
CO2  1  0  0 0  0 2  0 0       69    999    aq
Mg+2 0  0  0 0  1 0  0 2        9    999    aq
SiO2 0  0  0 0  0 2  1 0       72    999    aq
O2   0  0  0 0  0 2  0 0     2691    999   gas
H+   0  0  0 1  0 0  0 1        3     -7    aq

> species(c("quartz", "talc", "forsterite", "tremolite", "diopside", "wollastonite",
+     "monticellite", "merwinite"))

  HCl H2O Ca+2 CO2 Mg+2 SiO2 O2  H+ ispecies logact state         name
1   0   0    0   0    0    1  0   0     2005      0   cr1       quartz
2   0   4    0   0    3    4  0  -6     2030      0    cr         talc
3   0   2    0   0    2    1  0  -4     1920      0    cr   forsterite
4   0   8    2   0    5    8  0 -14     2032      0    cr    tremolite
5   0   2    1   0    1    2  0  -4     1891      0    cr     diopside
6   0   1    1   0    0    1  0  -2     2034      0    cr wollastonite
7   0   2    1   0    1    1  0  -4     1976      0    cr monticellite
8   0   4    3   0    1    2  0  -8     1972      0    cr    merwinite

> a <- affinity(`Mg+2` = c(-12, -4), `Ca+2` = c(-8, 0), T = 300, P = 1000)

affinity: temperature is 300 C
affinity: pressure is 1000 bar
energy.args: variable 1 is Mg+2 at 128 increments from -12 to -4
energy.args: variable 2 is Ca+2 at 128 increments from -8 to 0
subcrt: 16 species at 573.15 K and 1000 bar (wet)

> diagram(a)

diagram: immobile component is SiO2
diagram: conservation coefficients are 1 4 1 8 2 1 1 2
```



The 999's in the assignment of logarithms of activities of basis species could be any number – these settings do not affect the outcome of the calculation. This is so because 1) HCl, CO_2 and O_2 have zero stoichiometric coefficients in the species, 2) the activities of Ca^{+2} and Mg^{+2} correspond to the axes of the diagram, and their ranges are taken from the call to `affinity()`, and 3) SiO_2 is the immobile (conserved) component. Also note that "Mg+2" and "Ca+2" are not valid names of objects in R, but we can use them as names of arguments by putting them in quotation marks in the call to `affinity()`.

Here, the scales of the axes here depend on the pH setting. This calculation is therefore logically different from the formulation used in Ref. [5], where the axes are $\log\left(a_{\mathrm{Mg^{+2}}}/\sigma_{\mathrm{Mg^{+2}}}a^2_{\mathrm{H^+}}\right)$ and $\log\left(a_{\mathrm{Ca^{+2}}}/\sigma_{\mathrm{Ca^{+2}}}a^2_{\mathrm{H^+}}\right)$ (where σ is a function of the solvation of the subscripted species). However, the geometry of the stability fields in the diagram produced here is consistent with the previous work.

In just a few lines it's possible to make a wide variety of activity diagrams for organic and inorganic species. Try it for your favorite system!

9 Where to go from here

You can explore the package documentation through R's help system; just type `help.start()` at the command line and select CHNOSZ in the browser window that comes up. If you want to get an idea of the types of calculations available in CHNOSZ, run the examples in the help files for individual functions. A good one to try out might be `diagram()`; you can run all of the examples there with a single command:

```
> example(diagram)
```

Or you can use the following to run *all* of the examples provided in the documentation for the package. You will see a lot of text fly by on the screen, as well as a variety of plots. The examples will take about 5–10 minutes to run, depending on your machine.

```
> examples()
```

If you want to add to or modify the thermodynamic database, read the instructions at the top of the help page for `thermo`:

```
> help(thermo)
```

Have fun!

10 Document information

Revision history:

- 2010-09-30 Initial version
- 2011-08-15 Add `browse.refs()`; modifying database hint changed to `help(thermo)`

R session information:

```
> sessionInfo()

R version 2.13.1 (2011-07-08)
Platform: x86_64-slackware-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US       LC_NUMERIC=C         LC_TIME=en_US        LC_COLLATE=C
 [5] LC_MONETARY=C        LC_MESSAGES=en_US    LC_PAPER=en_US       LC_NAME=C
 [9] LC_ADDRESS=C         LC_TELEPHONE=C       LC_MEASUREMENT=en_US LC_IDENTIFICATION=C

attached base packages:
[1] tools     stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] CHNOSZ_0.9-7
```

References

[1] J. M. Dick. Calculation of the relative metastabilities of proteins using the CHNOSZ software package. *Geochem. Trans.*, 9:10, 2008. doi: 10.1186/1467-4866-9-10.

[2] J. M. Dick, D. E. LaRowe, and H. C. Helgeson. Temperature, pressure, and electrochemical constraints on protein speciation: Group additivity calculation of the standard molal thermodynamic properties of ionized unfolded proteins. *Biogeosciences*, 3(3):311 – 336, 2006. doi: 10.5194/bg-3-311-2006.

[3] J. W. Johnson, E. H. Oelkers, and H. C. Helgeson. SUPCRT92: A software package for calculating the standard molal thermodynamic properties of minerals, gases, aqueous species, and reactions from 1 to 5000 bar and 0 to 1000°C. *Comp. Geosci.*, 18(7):899 – 947, 1992. doi: 10.1016/0098-3004(92)90029-Q.

[4] G. M. Anderson. *Thermodynamics of Natural Systems*. Cambridge University Press, 2nd edition, 2005. URL `http://www.cambridge.org/0521847729`.

[5] T. S. Bowers, K. J. Jackson, and H. C. Helgeson. *Equilibrium Activity Diagrams for Coexisting Minerals and Aqueous Solutions at Pressures and Temperatures to 5 kb and 600° C*. Springer-Verlag, Heidelberg, 1984. URL `http://www.worldcat.org/oclc/11133620`.