Computational Identification of Candidate Nucleotide Cyclases in Higher Plants

Aloysius Wong and Chris Gehring

Chemical and Life Science and Engineering, King Abdullah University of Science and Technology, 23955-6900, Kingdom of Saudi Arabia

Summary

In higher plants guanylyl cyclases (GCs) and adenylyl cyclases (ACs) cannot be identified using BLAST homology searches based on annotated cyclic nucleotide cyclases (CNCs) of prokaryotes, lower eukaryotes or animals. The reason is that CNCs are often part of complex multi-functional proteins with different domain organizations and biological functions that are not conserved in higher plants. For this reason, we have developed CNC search strategies based on functionally conserved amino acids in the catalytic center of annotated and/or experimentally confirmed CNCs. Here we detail this method which has led to the identification of > 25 novel candidate CNCs in *Arabidopsis thaliana*, several of which have been experimentally confirmed *in vitro* and *in vivo.* We foresee that the application of this method can be used to identify many more members of the growing family of CNCs in higher plants.

Key Words: Cyclic nucleotide cyclase, Adenylyl cyclase, cAMP, Guanylyl cyclase, cGMP, Catalytic center, Motif search, Homology modeling, Basic Local Alignment Search Tool (BLAST), *Arabidopsis thaliana*.

Running Head: Identification of Plant Nucleotide Cyclases

1. Introduction

Adenylyl cyclases (ACs) and guanylyl cyclases (GCs) are cyclic nucleotide cyclases (CNCs) that catalyse the reaction from ATP respectively GTP to the messengers cAMP or cGMP. Particularly the role of cGMP in many plant responses is well documented and includes responses to light ***(1)***, hormones and signaling peptides ***(2-4)***, salt and drought stress ***(5,6)***, and ozone and pathogens ***(7,8)***. Cyclic nucleotides can also directly affect cellular ion homeostasis by gating an entire class of ion channels, the cyclic nucleotide gated channels (CNGCs) ***(9)***. Given the importance of the role of cyclic nucleotides it is not surprising that there is considerable interest in the enzymes that generate these molecules.

However, plant molecules with CNC activity are outside the detection limit of BLAST searches and other biochemical tools such as specific antibodies against CNCs from e.g. bacteria or animals because of the high level of divergence and complexity of CNCs which often combine two or more different domains ***(10,11)***. Examples of complex GCs with several domains are the plant leucine-rich

receptor kinases many of which contain extracellular ligand-binding domains and intracellular kinase and GC domains ***(4,12)***. We have therefore proposed and tested a search strategy that uses search motifs based on conserved amino acids in the catalytic center of experimentally tested CNCs ***(10,12,13)***. The residues include in position 1 the amino acid that does the hydrogen bonding with ATP or GTP, in position 3 the amino acid that confers substrate specificity and in position 14 the amino acid that stabilizes the transition state (ATP to cAMP or GTP to cGMP) and the C-terminal Mg^{2+}/Mn^{2+} binding site ***(14)*** (**Fig. 1**). Additional search conditions like the presence of a glycine-rich N-terminal domain or the presence of additional motifs such as an H-NOX motif ***(15)*** diagnostic for gas binding will add stringency to the candidate protein selection and allow for the identification of specific classes of functionally defined candidate CNCs. Confidence in identified candidate CNCs increases if firstly, structure modeling indicates that the catalytic center can assume a fold that is compatible with the functional requirements and secondly, if reciprocal BLAST with closely related species confirm that orthologous sequences also contain the conserved motifs and the catalytic centers.

A predictive 3D structure of the candidate CNCs can be constructed using homology modeling of candidate CNCs with known structures deposited in the Protein Data Bank (PDB) (http://www.rcsb.org/pdb/home/home.do). The key steps involved in protein homology modeling are template structure selection, the construction of protein models based on selected template, and the verification of these models ***(16)***. A structural model can provide important clues to the protein

function, which in this case is the CNC activity. Here we propose the use of the "Modeller" software ***(17)*** to construct 3D models for candidate CNCs. Orthologous sequences found in reciprocal BLAST provide additional information on the evolutionary conservation of these catalytic centers considered essential for CNC function and together with the predicted 3D models, they can further support the search for novel candidate CNCs in plants.

The method outlined here is particularly straightforward when working with *Arabidopsis thaliana* mainly because of the many on-line tools that are freely available in the public domain. However, in principle it can be applied successfully to many other species as long as a significant amount of sequence data of the species is available.

2. Materials

2. Homology modeling and search for orthologs

1. Download and install the "Modeller 9.10" software from the website http://salilab.org/modeller/download_installation.html.
2. The candidate example proteins modeled in Sections 3.2 and 3.3 are the brassinosteroid receptor AtBRI-GC (At4g3900), an annotated monooxigenase AtNOGC1 (At1g62580), and a diacylglycerol kinase AtDGK4 (At5g57690).
3. Download the respective protein crystal structures from the Protein Data Bank website at http://www.rcsb.org/pdb/home/home.do. The reference

crystal structures in Sections 3.2 and 3.3 are the catalytic domain of a eukaryotic guanylate cyclase (PDB entry: 3ET6), the bacterial nitric oxide sensor (PDB entry: 1XBN) and the *Escherichia coli* lipid kinase (*YegS*) (PDB entry: 2BON).

4. Download the "UCSF Chimera" software at http://www.cgl.ucsf.edu/chimera/download.html.

3. Methods

The method detailed here is designed to identify candidate CNCs in *Arabidopsis thaliana*, however ACs and GCs in other species can be inferred if the orthologs of these candidate CNCs also contain the search motif or relaxed search motifs.

3.1. Search for candidate CNCs in Arabidopsis

1. Download annotated CNC sequences from protein data repositories (e.g. NCBI (http://www.ncbi.nlm.nih.gov) or UniProtKB/Swiss-Prot (http://web.expasy.org/docs/swiss-prot_guideline.html) and select entries of the functional class of CNC of interest (*see* **Note 1**). Consider both entries from closely related and distantly related species (**Fig. 1A**).
2. Identify the catalytic center (or other domains of interest) of the annotated CNC and align them with an alignment program (e.g. "X" available at http://www.clustal.org/clustal2/) and curate the alignment by hand so that a motif can be built (**Fig. 1A**).

3. To build the search motif/search pattern by including all amino acid in the vertical alignment. Gaps of various lengths can be included and undefined amino acids are marked as “X” (*see* **Note 2**).

2. Once a search motif (pattern) has been built, open the TAIR web site (www.arabidopsis.org) and pull down the “Tools” menu, then go to the “Patmatch” function. In the “Patmatch” function choose peptide sequence of pattern and enter the search pattern. In the first instance the default options should be applied, searching “TAIR10” proteins with “no mismatch” allowed. The searches should begin with the most stringent motif and in subsequent searches the stringency can be relaxed e.g. by increasing the gaps or omitting residues (**Fig. 1B**) that may not be essential for catalysis (*see* **Note 3**).

3. Particularly given that the number of hits for a relaxed motif can be large, it is worth considering additional secondary search criteria to narrow the search. These may include e.g. a glycine-rich domain immediately N-terminal of the catalytic center and/or a pyrophosphate (PPi) binding motif that consists of an arginine (R) flanked by aliphatic amino acids (20-30 AA) N-terminal of the catalytic center.

4. A search with two or more motifs e.g. for the detection of candidate gas sensing CNC can be performed (*see* **Note 4**).

3.2. Search for ortholog sequences

1. Open the NCBI website (http://www.ncbi.nlm.nih.gov), insert the name of the desired protein in the search column and search against the protein database.
2. On the results page, click on the correct protein match then select “run BLAST” function at the right column.
3. On the BLAST page, select the “non-redundant protein sequences” option under the search set database and the “BLASTP” (protein-protein BLAST) option under the program selection. Then hit the BLAST button (*see* **Note 5**).
4. To increase the stringency of the search to identify a particular domain(s) within these search results (for example, the CNC and/or the H-NOX domains), repeat the BLAST search selecting the “phi-BLAST” (Pattern Hit Initiated BLAST) option in the program selection.
5. Insert the pre-defined CNC and/or H-NOX motifs, and hit BLAST (*see* **Note 6**). This will identify ortholog sequences that also harbor the predicted CNC and/or H-NOX domains (*see* **Note 7**).
6. Analyze the ortholog list and make rational inferences by, for example, considering the species of the orthologs and the evolutionary conservation of the predicted domain in the orthologs. For example, a phi-BLAST search with the H-NOX motif within the ortholog list (approximately 100 orthologs) of a full-length AtDGK4 (TAIR entry: At5g57690) BLAST search returns only eight candidates, seven of which are plant dicots. This suggests a highly conserved kinase catalytic center across species, but a

gas-binding domain that is unique only to plants and specifically the dicots. If this is true, then the plant DGKs (at least in the dicots) have evolved to be gas-sensing molecules.

3.3. Homology modeling

1. Open the NCBI website (http://www.ncbi.nlm.nih.gov), insert the name of the desired protein in the search column and perform a search against the protein database.
2. In the results page, choose the correct protein match and select “FASTA” function beneath the protein title. Save the FASTA file of the desired protein by clicking the “Send to” option on the right of the protein title (*see* **Note 8**).
3. Also in the previous BLAST result page, choose the correct protein match, then analyze the protein sequence by selecting the “run BLAST” function at the right column of the website.
4. Perform a BLAST search of the desired protein sequence against the “Protein Data Bank proteins” database and with the “BLASTP” (protein-protein BLAST) program selection. Hit the BLAST button and the website will return a list of known related crystal structures arranged by default in descending order of the “Expect value (E-value)”. One can also sort the list in the order of “max score”, “max identity” or “query coverage” by clicking on the respective title headers. In the case of CNC candidates, modeling is done against templates with the following criteria in

descending order of priority: 1) Sequence identity, 2) Sequence coverage, and 3) Max score (*see* **Note 9**).

5. Select the desired template structures from the previous BLAST search then go to the “Protein Data Bank” website (http://www.rcsb.org/pdb/home/home.do) and download the respective template PDB text files. To do this, insert the PDB IDs of the templates on the search bar and in the resulting page, pull down the “download files” function on the right column and select PDB file (text) to save the template protein structures (*see* **Note 10**).
6. Download the software “Modeler 9.10” and register for a license at http://salilab.org/modeller/download_installation.html (*see* **Note 11**).
7. Open the “Modeller” application file, run Scripts 1-5 for a complete modeling of the candidate protein. The “Modeller” scripts are written in the python programming language. Examples of scripts can be downloaded from http://salilab.org/modeller/tutorial/basic.html (*see* **Note 12**).
8. Follow the instructions detailed on the “Modeller” tutorial page at http://salilab.org/modeller/tutorial/basic.html to run the “Modeller” scripts. Script 1 instructs “Modeller” to search for template structures related to the protein of interest; script 2 allows the user to select one or more suitable templates; script 3 aligns protein of interest to the selected template; script 4 builds the models; while script 5 validates the quality of the constructed models.

9. View and assess the “best” model in “UCSF Chimera” or with other protein structural visualization software. The “best” model can be determined by assessing the molpdf, DOPE and GA341 scores from the log file of script 5. In the case of the candidate CNCs, additional model evaluation using the “Ramachandran” plot can be performed by uploading the PDB file of the “best” model at http://mordred.bioc.cam.ac.uk/~rapper/rampage.php (*see* **Note 13**).

10. These modeling procedures are for model building when only the amino acid sequence of the protein of interest is known. In reality, template structures and/or alignments may already been performed in another program. It is at the user’s discretion to skip one or more steps accordingly.

11. When assessing the 3D model, highlight the predicted functional domains (e.g. catalytic center of CNCs, H-NOX domain or ATP binding site) and analyze the structural properties such as the shape and conformation of the function domain, spatial and hydrophobic interactions, and the domain organizations within the molecule to determine functional compatibility. Also, assess these properties against known structures for further verifications. For example, based on the structural model of the candidate GC AtBRI-GC (TAIR entry: At4g3900), the functionally-assigned residues are organized in a manner that is similar to the reference crystal structure of a eukaryotic soluble GC (**Fig. 2A**).

12. Since many identified candidate CNCs have catalytic domains embedded within a kinase, the kinase region can also be modeled to evaluate the

structural features of the multifunctional protein. For example, the model for a candidate GC AtDGK4 (TAIR-annotated as a diacylglycerol kinase), reveals an ATP-binding site that sits within a cavity presumably ideal for catalytic function, and is in agreement with other known lipid kinase structures (**Fig. 2B**).

13. Similarly, based on the model, the predicted H-NOX domain of AtNOGC1 (TAIR entry: At1g62580) is compared to the reference NO-sensing crystal structure from bacteria (**Fig. 2C**). The arrangement of the functionally important residues suggests a plausible heme environment that can conceivably incorporate a porphyrin ring, hence providing the rational for heme-gas interactions (*see* **Note 14**).

4. Notes

1. Cyclic nucleotide binding proteins and CNC are a highly diverse group of proteins that has been functionally classified ***(18)***. It is therefore essential to study the classification and it is highly beneficial to look at the sequence logos (motifs) that have been proposed for the different classes of the nucleotide cyclase superfamily.
2. Since different search engines and search tools vary in the syntax they use, it is essential to consult the sample motif or search term for each program. The examples given (**Fig. 1**) use the syntax used by TAIR.

3. If a motif search does not return any hits, it is advisable to include chemically related amino acids into the search pattern. An example of such inclusions/substitutions of a chemically related amino acid is isoleucine (I in position 4) and leucine (L in position 9) (**Fig. 1A**). The latter has led to the discovery of AtGC1 (TAIR entry: At5g05930.1).
4. The Protein Information Resource PIR (http://pir.georgetown.edu/pirwww/search/pattern.shtml) also allows on-line pattern searches in UniProt and is particularly useful for a combined search with more than one motif/pattern. Note that the pattern should be entered in both orientation (A – B and B – A) and that the gap between the two motifs can be chosen. Note that the syntax is different from the one used in TAIR hence consult the instructions in “user-defined pattern”.
5. A BLAST with the full-length CNC will return sequences across species, hence providing information about the evolutionary diversity of the protein of interest. To locate specific domains within these results, the respective motifs must be included in the phi-BLAST function.
6. The phi-BLAST function is sensitive to only a specific set of pattern syntax rules that can be obtained at http://www.ncbi.nlm.nih.gov/blast/html/PHIsyntax.html.
7. The presence of CNC or H-NOX domains in the ortholog sequences provides a degree of confidence in the ability of the candidate proteins to perform their predicted functions. The type of species of the orthologs will

provide clues about the evolutionary conservation or diversification of these predicted domains.

8. The amino acid sequence of the candidate protein is required for subsequent modeling procedures. The downloaded FASTA file of the protein sequence must be converted to PIR database format (http://salilab.org/modeller/9v8/manual/node454.html), which is recognized by the “Modeller” software. For an extensive tutorial on how to use the “Modeller” program, see http://salilab.org/modeller/tutorial/basic.html.

9. For a detailed description about the BLAST scores, please refer to the Fall/Winter 2006/07 (Vol. 15, Issue 2) NCBI newsletter available online at http://www.ncbi.nlm.nih.gov/Web/Newsltr/V15N2/BLView.html. Templates may have high percentage identity but low sequence coverage to the protein of interest and vice versa. In principle, templates with identity percentage of >50% can generate good quality models. However, an identity percentage of 16-30% (depending on individual genomes) may be sufficient to construct reasonable models ***(16)***. If a particular domain of the protein is of high importance to the overall function of the protein, then it is advisable to select templates with higher identity percentage to the region of interest of the protein. However, the selection of suitable templates is at the user’s discretion.

10. The crystal structure of templates (in text file) is required for subsequent modeling procedures. Alternatively, templates alignment and selection can also be performed by running “Script 1” on “Modeller”. For a complete

tutorial on how to use the “Modeller” program, please see http://salilab.org/modeller/tutorial/basic.html.

11. The “Modeller” software is free. After installation, a registration using an institutional email address is required. For further download, installation and registration instructions please see http://salilab.org/modeller/download_installation.html.

12. For non-computer scientists/bio-infomaticians, the downloaded scripts can be modified and applied for most modeling applications. To do this, first change the script files extension from “.py” to “.txt”. Secondly, open the text script files and replace the names of the default protein and templates to user-specified protein file names. Then, save the edited script files in python format (.py), readable by “Modeller”.

13. In the log file for script 5, the “best” model is determined by the lowest Modpdf and DOPE scores or the highest GA431 score. The GA341 score ranges from 0.0 (worst) to 1.0 (native-like). The Modpdf and DOPE scores are not absolute measures as they only indicate relative model quality that is, they only rank models calculated from the same alignment. For assessment using the “Ramachandran” plot, a percentage of >90% of residues falling in the allowed region usually indicates good quality models.

14. The ability of AtBRI-GC to function as a GC has been proven experimentally ***(12)*** while *in vitro* evidence also confirmed AtNOGC1 to be both a GC and a gas-sensing molecule biased towards NO ***(15)***. The AtDGK4 has kinase catalytic domain that is highly conserved across

species and has ATP-binding site that reflects the consensus sequence of 'GXGG'.

5. References

1. Neuhaus, G., Bowler, C., Hiratsuka, K., Yamagata, H., Chua, N.H. (1997) Phytochrome-regulated repression of gene expression requires calcium and cGMP. *Embo J* 16:2554-2564.
2. Pharmawati, M., Billington, T., Gehring, C.A. (1998) Stomatal guard cell responses to kinetin and natriuretic peptides are cGMP dependent. *Cell Mol Life Sci* 54:272-276.
3. Gehring, C.A., Irving, H.R. (2003) Natriuretic peptides - a class of heterologous molecules in plants. *Int J Biochem Cell Biol* 35:1318-1322.
4. Kwezi, L., Ruzvidzo, O., Wheeler, J.I., Govender, K., Iacuone, S., Thompson, P.E., Gehring, C., Irving, H.R. (2011) The phytosulfokine (PSK) receptor is capable of guanylate cyclase activity and enabling cyclic GMP-dependent signaling in plants. *J Biol Chem* 286:22580-22588.
5. Maathuis, F.J., Sanders, D. (2001) Sodium uptake in Arabidopsis roots is regulated by cyclic nucleotides. *Plant Physiol* 127:1617-1625.
6. Donaldson, L., Ludidi, N., Knight, M.R., Gehring, C., Denby, K. (2004) Salt and osmotic stress cause rapid increases in Arabidopsis thaliana cGMP levels. *FEBS Lett* 569:317-320.
7. Pasqualini, S., Meier, S., Gehring, C., Madeo, L., Fornaciari, M., Romano, B., Ederli, L. (2009) Ozone and nitric oxide induce cGMP-dependent and -

independent transcription of defence genes in tobacco. *New Phytol* 181:860-870.

8. Qi, Z., Verma, R., Gehring, C., Yamaguchi, Y., Zhao, Y., Ryan, C.A., Berkowitz, G.A. (2010) Ca^{2+} signaling by plant *Arabidopsis thaliana* Pep peptides depends on AtPepR1, a receptor with guanylyl cyclase activity, and cGMP-activated Ca^{2+} channels. *P Natl Acad Sci USA* 107:21193-21198.
9. Leng, Q., Mercier, R.W., Yao, W., Berkowitz, G.A. (1999) Cloning and first functional characterization of a plant cyclic nucleotide-gated cation channel. *Plant Physiol* 121:753-761.
10. Ludidi, N., Gehring, C. (2003) Identification of a novel protein with guanylyl cyclase activity in Arabidopsis thaliana. *J Biol Chem* 278:6490-6494
11. Meier, S., Seoighe, C., Kwezi, L., Irving, H., Gehring, C. (2007) Plant nucleotide cyclases: an increasingly complex and growing family. *Plant Signal Behav* 2:536-539.
12. Kwezi, L., Meier, S., Mungur, L., Ruzvidzo, O., Irving, H., Gehring, C. (2007) The Arabidopsis thaliana brassinosteroid receptor (AtBRI1) contains a domain that functions as a guanylyl cyclase in vitro. *PloS One* 2:e449.
13. Gehring, C. (2010) Adenyl cyclases and cAMP in plant signaling - past and present. *Cell Commun Signal* 8:15.
14. Liu, y., Ruoho, A., Rao, V., Hurley, J. (1997) Catalytic mechanisms of the adenyl and guanylyl cyclases: Modelling and mutational analysis. *P Natl Acad Sci USA* 94:13414-13419.

15. Mulaudzi, T., Ludidi, N., Ruzvidzo, O., Morse, M., Hendricks, N., Iwuoha, E., Gehring, C. (2011) Identification of a novel Arabidopsis thaliana nitric oxide-binding molecule with guanylate cyclase activity in vitro. *FEBS Lett* 585:2693-2697.

16. Krieger, E., Nabuurs, S.B., Vriend, G. (2005) Homology Modeling. Structural bioinformatics. John Wiley & Sons, Inc., New York, pp. 509-523.

17. Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U., Sali, A. (2001) Comparative protein structure modeling using MODELLER. Current Protocols in Protein Science. John Wiley & Sons, Inc., New York, Chapter 2:Unit 2.9.

18. McCue, L., McDonough, K., Lawrence, C. (2000) Functional classification of cNMP-binding proteins and nucleotide cyclases with implications for novel regulatory pathways in Mycobacterium tuberculosis. *Genome Res* 10:204-219.

Figure legends:

Figure 1 Example of an alignment of CNC catalytic centres and the building of search motifs for candidate CNC. (**A**) Edited ClustalX alignment of catalytic centres from annotated GC from different species. In the deduced 14 amino acid motif the substitutions are in square brackets ([ ]), "X" stands for any amino acid and the gap size is marked in curly brackets ({ }). The underlined amino acids have been added to the motif because of their chemical similarity to the amino acid at this position. The amino acids in red are functionally annotated ***(10)***. The open arrow ( ) signifies the glutamic acid (E) implicated in Mg^{2+} respectively Mn^{2+} binding (not included in the original search motif). (**B**) Modifications of the motif for the discovery of candidate AC. The residues in position three confer substrate specificity and have been changed to D or E (in blue) to recognize ATP rather than GTP. The Mg^{2+} respectively Mn^{2+} binding residues are marked green.

Figure 2 Examples of protein models constructed using the "Modeller" software and how they compare to their respective reference crystal structures. (**A**) The AtBRI-GC catalytic center (left) was modeled against the *Pseudomonas syringae* AvrPtoB (PDB entry: 3TL8) and compared to the reference structure (right). The GC domain is highlighted in yellow and the functionally-assigned residues of the GC catalytic center are detailed. The amino acid residue in position one (R/K/S) of the GC motif does the hydrogen bonding with GTP; the residue in position three (C/G/T/H) confers substrate specificity; while the

position 14 amino acid stabilizes the transition state of GTP to cGMP. The 16^{th}/17^{th} residue (D/E) at the C-terminal of the 14-amino acid GC motif suggests binding to Mg^{2+}/Mn^{2+}. (**B**) The AtDGK4 kinase domain (left) was modeled against the crystal structure of *Salmonella typhimurium YegS* (PDB entry: 2P1R). The ATP-binding site highlighted in cyan and forms part of a helical coil that is buried in a cavity (insets) which is in agreement to that of the reference structure (right). (**C**) The AtNOGC1 gas-binding domain (left) was modeled against the crystal structure of a bacterial flavin-containing monoxygenase (PDB entry: 2VQ7). Functionally important residues are represented and their interactions with NADP are shown in the reference structure (right). All images are created using the "UCSF Chimera" software. Note that the AtBRI-GC has GC activity *in vitro* ***(12)***; the AtDGK4 is annotated as diacylglycerol kinase (TAIR) and has an ATP-binding site that reflects the 'GXGG' nucleotide-binding consensus sequence and the AtNOGC1 has GC and gas-binding activity *in vitro*.