Journal of Biotechnology 135 (2008) 340–350

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

The GlxR regulon of the amino acid producer *Corynebacterium glutamicum*: *In silico* and *in vitro* detection of DNA binding sites of a global transcription regulator

Thomas A. Kohl[a,b], Jan Baumbach[b], Britta Jungwirth[a], Alfred Pühler[c], Andreas Tauch[a,*]

[a] *Institut für Genomforschung und Systembiologie, Centrum für Biotechnologie, Universität Bielefeld, Universitätsstraße 25, D-33615 Bielefeld, Germany*
[b] *International NRW Graduate School in Bioinformatics and Genome Research, Centrum für Biotechnologie, Universität Bielefeld, Universitätsstraße 25, D-33615 Bielefeld, Germany*
[c] *Lehrstuhl für Genetik, Fakultät für Biologie, Universität Bielefeld, Universitätsstraße 25, D-33615 Bielefeld, Germany*

ARTICLE INFO

Article history:
Received 31 January 2008
Received in revised form 20 May 2008
Accepted 23 May 2008

Keywords:
Corynebacterium glutamicum
Transcriptional regulatory network
Motif discovery
Transcription factor binding site
Global regulator
GlxR

ABSTRACT

The *glxR* (*cg0350*) gene of *Corynebacterium glutamicum* ATCC 13032 encodes a DNA-binding trancription regulator of the CRP/FNR protein family. Five genomic DNA regions known to be bound by GlxR provided the seed information for DNA binding site discovery by expectation maximization and Gibbs sampling approaches. The detection of additional motifs in the genome sequence of *C. glutamicum* was performed with a position weight matrix and a profile hidden Markov model, both deduced from the initial motif discovery. A combined iterative search for GlxR binding sites revealed 201 potential operator sequences. The interaction of purified GlxR protein with 51 selected binding sites was demonstrated *in vitro* by performing electrophoretic mobility shift assays with double-stranded 40-mer oligonucleotides. Considering potential operon structures and the genomic organization of *C. glutamicum*, the expression of 53 transcription units comprising 96 genes may be controlled directly by GlxR. The DNA binding site of GlxR is apparently specified by the consensus sequence TGTGANNTANNTCACA. Integration of the data into the transcriptional regulatory network model of *C. glutamicum* revealed a high connectivity of the deduced regulatory interactions and suggested that GlxR controls at least (i) sugar uptake, glycolysis, and gluconeogenesis, (ii) acetate, lactate, gluconate, and ethanol metabolism, (iii) aromatic compound degradation, (iv) aerobic and anaerobic respiration, (v) glutamate uptake and nitrogen assimilation, (vi) fatty acid biosynthesis, (vii) deoxyribonucleotide biosynthesis, (viii) the cellular stress response, and (ix) resuscitation.

1. Introduction

Corynebacterium glutamicum is a well-studied, non-pathogenic soil bacterium of profound industrial importance, especially for the biotechnological production of amino acids. It was first isolated by S. Udaka and named according to its ability to accumulate significant amounts of glutamic acid in the growth medium (Udaka, 1960). Since the discovery of *C. glutamicum*, so-called high-performance strains have been developed for the fermentative production of the flavor enhancer monosodium glutamate and the feed additive L-lysine. The total annual production of L-amino acids has been estimated at two million tons worldwide, which are predominantly produced by engineered *C. glutamicum* and *Escherichia coli* strains. In addition to this traditional application, there have been proposals to use *C. glutamicum* under oxygen deprivation for the production of ethanol, lactate, and succinate (Inui et al., 2004a,b).

Due to the importance of *C. glutamicum* in industrial biotechnology, the genomes of two strains have been sequenced so far (Kalinowski et al., 2003; Yukawa et al., 2007). While classical approaches for strain development relied on iterative mutagenesis and screening procedures, the development of genomic, proteomic, and metabolomic technologies provided novel strategies for rational strain design by using metabolic and regulatory models of cellular sub-systems. A fundamental part of such *in silico* models is the transcriptional regulatory network, basic components of which are regulatory proteins that control the expression of target genes by binding to cognate operator sequences. An important feature of all biological networks characterized so far is their hierarchical structure, with a certain number of highly connected nodes serving as hubs for the network (Babu et al., 2006). The current knowledge about transcription regulators of *C. glutamicum* has been incorporated into a comprehensive network model (Brinkrolf et al., 2007) and has been integrated into the specialized database

* Corresponding author. Tel.: +49 521 106 5605; fax: +49 521 106 5626.
E-mail address: Andreas.Tauch@Genetik.Uni-Bielefeld.DE (A. Tauch).

0168-1656/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2008.05.011

CoryneRegNet (Baumbach, 2007). However, the size of the deduced *C. glutamicum* network was small (411 regulations) when compared with that of *E. coli* (2237 regulations; Gama-Castro et al., 2008), and the general connectivity was low, indicating the lack of data about regulatory hubs, so-called "global regulators".

A potential candidate for a regulatory hub is the GlxR protein (Cg0350) that has been characterized as cAMP-binding transcription regulator of the CRP/FNR protein family, showing homology to the global regulator CRP from *E. coli* (Kim et al., 2004). In an early study, the Cg0350 protein was suggested to regulate expression of the *dtsR1* (*accD1*) gene, and was therefore named *drp* (*dtsR1* regulating protein) (Hirano et al., 2001; Kimura, 2002). Later Cg0350 was characterized as regulator of the glyoxylate bypass genes *aceA* and *aceB*, and designated GlxR (Kim et al., 2004). The cAMP-dependent binding of GlxR has been demonstrated for five genomic DNA fragments located upstream of genes involved in the glyoxylate bypass, gluconate metabolism, glycolysis, and gluconeogenesis: *aceB* (Kim et al., 2004), *gntP* (Letek et al., 2006), *gntK* (Letek et al., 2006), *gapA* (Han et al., 2007), and *pck* (Han et al., 2007). The DNA binding site of GlxR has not yet been determined experimentally, but it was suggested to be similar to the consensus sequence of the *E. coli* CRP binding sites (Letek et al., 2006). Very recently, the construction of a *glxR* mutant has been stated, which exhibited growth defects and differential expression of more than one hundred proteins, although no supporting data was presented (Moon et al., 2007).

In this study, we applied a position weight matrix (PWM) and a profile hidden Markov model (HMM) deduced from sequences known to be bound by GlxR to detect further binding sites in the genome sequence of the wild-type strain *C. glutamicum* ATCC 13032. The interaction of purified GlxR protein with 40-mer oligonucleotides representing 51 predicted binding sites was demonstrated *in vitro* by electrophoretic mobility shift assays. The resulting data were integrated into the transcriptional regulatory network of *C. glutamicum*, indicating that GlxR represents the first regulatory hub in the network model.

2. Material and methods

2.1. Extraction of genomic sequences for motif discovery

An interface to the GenDB system (Meyer et al., 2003) was implemented to make use of the manually curated annotation of the *C. glutamicum* ATCC 13032 genome sequence (Kalinowski et al., 2003). Based on the distribution of known regulator binding sites in the genome sequence of *C. glutamicum* (Baumbach, 2007), upstream regions were defined as genomic sequences ranging from +20 bp relative to the translation start up to the next coding sequence (CDS), with a maximum size of 620 bp. For each annotated CDS, the upstream sequence was extracted for further analysis. CDS smaller than 150 bp were ignored when determining the upstream sequences. Furthermore, upstream regions shorter than 40 bp were excluded from further analysis. The resulting set of genomic sequences was used as background data set for motif discovery and as search space for the detection of GlxR binding sites.

2.2. Bioinformatics motif discovery

Five genomic DNA sequences of *C. glutamicum* for which binding of GlxR has been demonstrated experimentally (Kim et al., 2004; Letek et al., 2006; Han et al., 2007) were used as a training set. Motif discovery was performed by expectation maximization (EM) with MEME (Bailey and Elkan, 1994) and by Gibbs sampling using BioProspector (Liu et al., 2001). According to the operator consensus sequence of the CRP regulator from *E. coli* (Zheng et al., 2004), parameters for MEME were chosen to detect a 16 bp palindromic motif with five (minimum) to ten (maximum) occurrences in input data, with a maximum of 200 EM iterations. BioProspector was configured to search for a palindromic motif consisting of two blocks of 5 bp, separated by 6 bp, also with 200 iterations. A set of five potential GlxR binding sites confirmed by wet-lab analysis was used to model the binding motif as PWM by applying Perl modules (Lenhard and Wasserman, 2002) and a calibration for base frequencies in the aforementioned set of genomic sequences. In parallel, a HMM for the binding motif was built and calibrated with the programs *hmmbuild* and *hmmcalibrate* of the HMMER package (Eddy, 1998).

2.3. In silico detection of GlxR binding sites

PoSSuMsearch (Beckstette et al., 2006) and *hmmsearch* (Eddy, 1998) were applied to search the extracted genomic sequences of *C. glutamicum* independently of each other. An automated pipeline was implemented to execute programs and to combine their individual outputs. As *hmmsearch* does not support searching in both directions of a DNA double-strand, scanning the reverse complementary DNA sequences was implemented additionally. *PoSSuMsearch* was configured for lazy probability evaluation (Beckstette et al., 2006). For both *PoSSuMsearch* and *hmmsearch*, E-value thresholds were chosen in such a way that all previously confirmed binding sites were detected. Hits found by both tools were selected for further analysis. Additional information was used for manual curation of the search results, including known transcription start sites, functional annotation of target genes (Kalinowski et al., 2003), and operon predictions (Price et al., 2005).

2.4. Purification of His-tagged GlxR protein

For protein purification, we used plasmid pETCRP encoding the *C. glutamicum* GlxR protein fused to an amino-terminal hexahistidyl-tag (Letek et al., 2006). *E. coli* cells were grown at 37 °C in LB medium supplemented with 50 μg ml^{-1} kanamycin and set to 17 °C after induction with 1 mM IPTG. His-tagged GlxR protein was purified after cell rupturing using Protino Ni-TED 1000 columns (Macherey-Nagel). The purified protein was visualized by SDS-PAGE, and purification was verified by MALDI-TOF mass spectrometry.

2.5. Electrophoretic mobility shift assays (EMSAs)

Cy3-labeled, double-stranded DNA oligomers (40-mers) were used for EMSAs. The selected 40-mers consist of the 16 bp motif instance flanked on both sides by 12 bp of native genomic sequence. In the cases of the 40-mers for *gntK* and *gapA*, the oligonucleotide sequences were modified at the 5′ end of the genomic flanking sequences to distort extended inverted repeats. To test for specific interaction of GlxR with proposed binding sites, mutated sequences were employed by introducing transitions. First, the conserved pentamers at either end of the 16 bp motif were mutated. To complement this analysis, oligonucleotides with mutated 12 bp flanking sequences were generated. DNA band shift assays contained 0.1 pmol labeled 40-mer, 60 pmol purified GlxR, 200 μM cAMP, 13% (v/v) glycerin, 10 mM Na_2HPO_4, 137 mM NaCl, 2.7 mM KCl, and 0.05 μg herring sperm DNA in a 20 μl volume. The assays were incubated at 30 °C for 20 min, separated on 2% agarose gels, and visualized using a Typhoon 8600 imager (Amersham Biosciences).