Modulation of EGFR gene transcription by secondary structures, a polymorphic repetitive sequence and mutations - a link between genetics and epigenetics

F. Gebhardt, H. Bürger and B. Brandt

1Institute for Clinical Chemistry and 2Gerhard-Domagk-Institute for Pathology, University of Münster, Germany

Summary. The epidermal growth factor receptor (EGFR) plays a crucial role in growth, differentiation and motility of normal as well as tumor cells. The transduction of extracellular signals to the cytoplasm via the receptor not only depends on ligand binding, but is also determined by the receptor density on the cell surface. Therefore, in terms of cancer diagnosis and therapeutic approaches targeting EGFR it is decisive to know how the expression level of EGFR is controlled. We found that transcription activity declines with increasing numbers of CA dinucleotides of a highly polymorphic CA repeat in the first intron epidermal growth factor receptor gene. In vivo data from cultured cell lines support these findings, although other regulation mechanisms can compensate this effect. In addition, we showed that RNA elongation terminates at a site closely downstream of the simple sequence repeat (SSR) and that there are two separate major transcription start sites. Model calculations for the helical DNA conformation revealed a high bendability in the EGFR polymorphic region, especially if the CA stretch is extended. These data suggest that the CA-SSR can act like a joint bringing the promoter in proximity to a putative repressor protein bound downstream of the CA-SSR. The data suggest that this polymorphism is a marker for cancer linking genetic and epigenetic risk. Furthermore in breast cancer, heterozygous tumours with short CA-SSR showed an elevated EGFR-expression in contrast to tumours with longer CA-SSR. Tumours with loss of heterozygosity in intron 1 of egfr revealed an increased EGFR expression if the longer allele was lost. Moreover, deceased egfr gene dosages were significantly correlated to poor prognosis in breast cancer.

Key words: EGFR, Gene transcription, Gene regulation, CA repeats, Loss of heterozygosity

The ERBB gene family

The c-erbB family of proto-oncogenes plays an important role in cell growth, differentiation and motility of normal as well as tumor cells. The transduction of extracellular signals to the cytoplasm via c-erbB receptors does not only depend on the ligands binding to the extracellular domain, but is also determined by concentrations of the different receptors on the cell surface. Therefore, it is crucial to know how the expression levels of c-erbB genes are controlled. Besides post-translational mechanisms, alteration of gene copy numbers or of mRNA levels are two major ways in which the expression of genes can be regulated.

The v-erbB oncogene is carried by the retroviruses ALV (avian leukosis virus), ASV (avian sarcoma virus), and AEV (avian erythroblastosis virus). In the AEV-ES4 strain, it is accompanied by the v-erbA gene (Rothe Meyer and Engelbreth-Holm, 1933). The two genes are functionally not related. The v-erbA product is a high affinity receptor for the thyroid hormone triiodothyronine (T3) while v-erbB codes for a 68 kDa glycoprotein with constitutive tyrosine kinase activity. v-erbB causes rapid induction of erythroblastosis and sarcomas after intramuscular injection of the carrying virus (Hesketh, 1994). The first human cellular homologue of the v-erbB gene to be identified was the epidermal growth factor receptor (egfr/c-erbB-1) gene. The viral oncoprotein is almost identical to the transmembrane and cytoplasmatic domain of EGFR. Unlike v-erbB, EGFR carries an extracellular domain and is activated after ligand binding. It was initially localized to chromosome 7p13-q22 (Kondo and Shimizu, 1983) and later mapped more precisely to p13-p12. The egfr locus covers 110 kbp and 26 exons (Haley et al., 1987). The cDNA consists of 3630 bp (Ullrich et al., 1984). To date, three other members of the erbB family of tyrosine kinase receptors have been identified: c-erbB-2 (HER-2/neu), c-erbB-3 (HER-3), and c-erbB-4 (HER-4). They all share broad homologies with egfr/c-erbB-1. Before the human c-erbB-2 proto-oncogene was
known, the rat homologue neu was identified in studies with NIH 3T3 cells that were transformed with DNA from ethylnitrosourea-induced rat neuroblastomas (Padhy et al., 1982). After that, c-erbB-2 was found on the human chromosome 17q21-q22 by screening genomic and cDNA libraries (Coussens et al., 1985; King et al., 1985a; Semba et al., 1985). The third member of the gene family, c-erbB-3 was found on chromosome 12q23 (Kraus et al., 1989). It is translated to a 6.2 kb full length mRNA and an alternatively spliced truncated transcript of 1.4 kb encoding a secreted receptor (Kato et al., 1993). The c-erbB-4 gene codes for two mRNA species of 6.0 and >15 kb (PloSWman et al., 1993) and was localized to chromosome 2q33.3-q34 (Zimonjic et al., 1995).

The epidermal growth factor receptor (EGFR)

The EGFR receptor stimulates cell growth and differentiation after binding of specific ligands (Savage and Cohen, 1972; Carpenter, 1987). It was named after the epidermal growth factor, which was discovered as the first EGFR ligand. Meanwhile, additional polypeptides like transforming growth factor α (TGF-α) and amphiregulin were identified as EGFR binding ligands (Laurence and Gusterson, 1990; Johnson et al., 1993). EGFR is a 170 kDa plasma membrane-bound receptor with intrinsic tyrosine kinase activity in the intracellular domain. Further, there is a transmembrane unit and an extracellular domain capable of ligand binding. After binding of a ligand, the receptor instantly autophosphorylates and dimerizes with another erbB receptor. Subsequently, the tyrosine kinase is activated before the protein is finally internalized and digested. Activation of the kinase domain leads to phosphorylation of important intracellular proteins that transduce the signal that results in proliferation or differentiation.

It has been found to be expressed in many normal and malignant cell types (Adamson and Rees, 1981). Overexpression of EGFR alone is sufficient to transform NIH 3T3 cells in an EGFR-dependent manner (DiFiore et al., 1987; Velu et al., 1987). In most cell types, EGFR is found in amounts varying from 2x10^4 to 2x10^5 receptors per cell. Human placenta expresses it at high levels (Carson et al., 1983; O’Keefe et al., 1984), oral mucosa moderately (Beenken et al., 1994), while hematopoietic cells reveal no detectable amounts.

Overexpression of EGFR has been described for many tumour tissues and is mostly associated with poor prognosis (Gullick et al., 1986; Gasparini et al., 1994; Klijn et al., 1994; Benz et al., 1995; Todd and Wong, 1999). Several studies showed a positive correlation of increased amounts of the receptor with shortened survival and failure of endocrine therapy in breast cancer (Boluffer et al., 1990). On this clinical background, understanding novel transcriptional regulation mechanisms that regulate EGFR proto-oncogene expression is important. Moreover, the knowledge of secondary structures, simple repetitive sequences like CA repeats (CA-SSR) and mutation function in relation to negative or positive enhancers provides new insights into individually different gene expression and the linkage of inherited polymorphisms to cancer.

EGFR gene transcription in normal and cancer cells

EGFR gene transcription results in at least two different mRNA species of 5.8 and 10.5 knt, respectively. These transcripts are found in all EGFR-expressing tissues. Injection into Xenopus oocytes demonstrated that both were translated to the 170 kDa EGFR receptors, but function of these two mRNAs is not yet clear (Simmen et al., 1984). Another e→r mRNA of only 1.8 kb was identified in normal human placenta tissue. It codes for the extracellular domain only and is produced via alternative RNA processing. Ilkis et al. (1995) suggested that after translation of the truncated message, the ligand binding protein can form heterodimers with full-length EGFR receptors. These heterodimers with complete or internally deleted receptors (EGFRvIII) were not capable of tyrosine kinase activation (Basu et al., 1989; O’Rourke et al., 1998). This may be a way of signal transduction regulation by suppressing the formation of active EGFR homodimers or heterodimers with other members of the erbB family.

The egfr gene promoter is GC-rich and contains no consensus sequences like TATA or CAAT boxes. Transcription starts at multiple initiation sites within the promoter region (Ishii et al., 1985). Further, three sequences with enhancer activity were found within the 5’ region of the gene. Two of them, located upstream of the promoter (-1409 to -1109) and downstream in intron 1 (+1788 to +2318) close to a polymorphic region showed cooperative function: the downstream enhancer functioned only in presence of the upstream element (Fig. 1; Maekawa et al., 1989). The third enhancer was found in direct proximity to the promoter (Fig. 1;
There are several transcription factor binding sites within the 5′-region of the egfr gene. Sp1 seems to play a major role in basal transcription regulation of the receptor gene (Kageyama et al., 1988a,b). ETF1 specifically stimulates transcription in vitro from the egfr and other promoters lacking a TATA box (Kageyama et al., 1989). TCF binds to a region with several repeats of the sequence TCC, found in the egfr promoter region (Johnson et al., 1988). The transcriptional repressor GCF binds at two sites in the egfr gene, that partially overlap with the ETF1 recognition site (Kageyama et al., 1989). Another protein of 128 kDa that binds to a repressor site about 900 bp upstream of the ATG start codon was found by Hou et al. (1994). Upregulated positive transcription factors or inactivated transcriptional repressors can lead to elevated expression (Kitada et al., 1992). In fact, expression of the receptor is mainly regulated at the level of gene transcription (Xu et al., 1984; Merlino et al., 1985).

The egfr mRNA production can be stimulated directly or indirectly by treatment of cells with EGF, phorbol, G-1,2,5-cAMP, dexamethasone, thyroid hormone, retinoic acids, interferon α, and wild type p53. In overexpressing cells the effect can be reversed, thus causing a suppression of the egfr promoter (Clark et al., 1985; Hudson et al., 1989,1990; Hamburger and Pinnamaneni, 1991; Xu et al., 1993; Deb et al., 1994; Subler et al., 1994; Grandis et al., 1996).

**Regulation of EGFR gene transcription: dependency on secondary structures and a polymorphic repetitive sequence**

Transcription around the exon 1 was found to be elevated 8-fold compared to exons 2 to 26. This observation was explained by a partial block of RNA elongation in intron 1, possibly caused by a stem loop structure of the DNA (Haley and Waterfield, 1991). Studies on in vitro and in vivo transcription of the egfr 5′-region by nuclear in vitro transcription (run-off) and RNA hybridization (ribonuclease protection assay) confirmed a transcription termination site in a very AT-rich region at +1900 in intron 1 (Gebhardt et al., 1997). Further evidence for a possible stem loop structure at position +2020 came from the detection of a shorter band derived from cell line A431 (Gebhardt et al., 1997). By S1 nuclease digestion three secondary structures in the egfr 5′-region could be probed: a weakly base paired poly-A stretch upstream the CA repeat polymorphic sequence which is also digested due to base mismatch caused by Taq amplification and a downstream triplex forming structure in the promoter region (Fig. 2). Additionally, an anti-sense RNA starting downstream of exon 1 and presumably causing suppression of translation was detected (Haley and Waterfield, 1991).

In a study of the acetyl-CoA carboxylase gene, it was demonstrated (Tae et al., 1994), that a non-polymorphic sequence of 28 CA-SSRs within the promoter region can repress the activity of one of the two promoters by 70%. However, in this special case mediation of promoter inhibition seems to require a CAAT box. Furthermore, it has been shown that the first intron of several genes including egfr has an important regulatory function (Bornstein et al., 1988; Franklin et al., 1991; Sica et al., 1992; Chrysogoles, 1993). A polymorphic simple sequence repeat with 14 to 21 CA dinucleotides and a heterozygosity of 72% in a Caucasian reference pedigree (Chi et al., 1992) was revealed close to the downstream enhancer element. In the case of EGFR a similar effect is observed although there is no CAAT box in the promoter region and the CA-SSR is located more than 1000 bp downstream of

**Fig. 2.** S1 nuclease digested PCR product from the 5′-region of the egfr gene. A. Separation of fragments on a agarose gel. B. Diagram of the 5′-region with the triplex forming site, the CA SSR and a poly-A-stretch. Fragment length are given in basepairs.
the promoter. Differences in the number of CA-SSRs in the egrf intron 1 show different levels of transcription modulation. EGFR transcription activity in vitro declines with increasing numbers of CA repeats in intron 1 (Fig. 3). In addition, it could be demonstrated by ribonuclease protection assays that transcription in vitro with A431 nuclear extracts terminates at a site near the polymorphic region in intron 1. These results suggest a dual function of the polymorphic region: First, an indirect effect that enhances or represses transcription in vitro up to fivefold depending on the number of CA repeats and second, a block of RNA elongation unaffected by the length of the CA-SSR. The observed effect in vitro has also important in vivo on the protein expression level. Allele dependent modulation of EGFR transcription can be observed in carcinoma cell lines in vivo, but not surprisingly, there are other regulation mechanisms that can compensate it (Gebhardt et al., 1999). Upregulation by the action of transcription factors could easily overcome this more basal effect. Bending of DNA in a sequence-dependent manner has an important function in many biological events like DNA replication, site-specific recombination and transcription. Helical conformation analysis on the basis of the CA-SSR and flanking sequences provide evidence for a possible involvement of the EGFR polymorphic region in conformational changes after transcription factor binding: The intrinsic DNA curvature propensity of the poly-CA stretch, a measure for helical asymmetry frequently associated with a rigid conformation, is remarkably low while the bendability is prominently elevated. The longer the CA stretch, the longer becomes the highly bendable section, too. This could enable a protein to bind preferentially to a sequence neighboring a long CA-SSR (Fig. 4). Especially, if the factor requires the DNA to bend back and form a loop, a flexible segment like the CA-SSR can act like a joint. Loop structures are responsible for the effect of distant enhancer elements in transcription, mediating activation of the promoter over thousands of basepairs. Distant negatively regulatory sequences in enhancer elements (“negative enhancers”, repressor elements) have also been described and an equivalent mechanism seems very likely. In the case of EGFR, action of a repressor protein that inhibits transcription of DNA molecules with preferably higher numbers of CA repeats could explain our findings. The highly bendable joint sequence is located between a downstream enhancer element with several binding sites and the promoter region. A loop structure, that brings together these regulatory elements by the action of one or more DNA binding proteins may be favored if the polymorphic stretch is prolonged. In this way, a transcriptional repressor protein that binds downstream of the CA-SSR could serve as a mediator of allele-dependent inhibition in EGFR transcription. This is a novel transcriptional regulation mechanism that involves the action of a presumed repressor dependent on the length of a CA-SSR and its influence on DNA flexibility.

Modified regulation of EGFR gene transcription due to mutations in the 5'-region of the gene

Mutations of the egrf gene have been described for several tumour types and also in premalignant lesions (Ullrich et al., 1984; King et al., 1985b; Merlino et al., 1985; Werkmeister et al., 1999). Among them, gene amplification (Stark and Wahl, 1984) is most frequently observed (Brison, 1993). Especially gliomas and other tumours of the central nervous system often reveal egrf amplifications, particularly in earlier tumour stages (Sugawa et al., 1990; Walton et al., 1990; Yamazaki et al., 1990; Ekstrand et al., 1992; Weng et al., 1992). Usually, egrf gene amplification leads to overexpression of the gene.

In glioblastomas, lung and breast carcinomas, egrf genes with an extended in-frame deletion from amino acid 6 to 273 (exons 2-7) were found (Moscatello et al.,

---

**Fig. 3.** Relative transcription activity of the 5'-region of the egrf gene in vitro.

**Fig. 4.** Bendability of the 5'-region of the egrf gene determined by the CA-SSR in intron 1 and potential binding of a transcription modulating protein.
Overexpression of EGFR is not always due to amplified genes. For breast cancer the amplification rate determined by Southern blotting are reported with 0-14% (Ro et al., 1988). Applying the more precise and sensitive method of double-differential PCR revealed amplifications with a gene copy number of greater than 3 for one fifth of the patients using primers in the 5’-region at the exon1-intron1-boundary of the egfr gene (Brandt et al. 1995). Also remarkable is that another fifth of the patients presented with significantly decreased levels of gene copy number of less than 0.4 (Brandt et al. 1995).

As mentioned before, all known regulatory sequences of the egfr gene are located within the 5’-flanking region and intron 1. Elevated levels of intratumoural EGFR concentrations have been measured in those tumours which showed either elevated or decreased gene dosages for egfr at the exon1-intron1-boundary (Fig. 5A; Brandt et al. 1997). Moreover, both subgroups of patients with deviated egfr gene dosages in combination with high EGFR concentrations in the tumour can be confined as a high risk subgroup for early onset of tumour progression (Fig. 5B; Brandt et al. 1997).

In contrast, loss of heterozygosity at the CA-SSR was not correlated with a higher EGFR expression. The samples with a loss of the longer allele showed a higher EGFR content in contrast to the tumours with a remaining longer one. Loss of the 16 CA repeat allele showed a tentative lower intratumourous protein concentration than loss of the longer allele. Breast cancer patients with a prognostic unfavourable EGFR overexpression showed almost exclusively a short allele with 16-18 CA repeats. In those cases the loss of the longer allele was frequently observed.

Those deletions are restricted to short sequences due to the fact that several studies showed losses or isolated gains of chromosome 7p in less than 5% of all tumours investigated (Kallioniemi et al., 1994; Ried et al., 1995; Burger et al., 2000). Like in other epithelial neoplasms an altered EGFR expression therefore is not reflected on the cytogenetic level (Moch et al., 1998) and microdeletions in the egfr gene play a much more
important role for gene expression. Furthermore the avoidance of deletions in this polymorphic CA-SSR sequence in the human genome could be a target for cancer prevention (Brandt et al., 1999). An association between reduced EGFR expression at the time of carcinogen exposure and a higher resistance towards development of breast cancer could be derived from experiments in rats (Brown et al., 1998). The hypothetical combination of inherited CA-SSRs with different length and their direct influence on EGFR expression and further losses of CA-SSRs due to chemical carcinogenesis might therefore bridge the gap between different genetic and epigenetic factors in the carcinogenesis at least of invasive breast cancer.

Acknowledgements. This article is dedicated to Prof. Fritz Anders, Gießen, on the occasion of his eighties birthday.

References


egfr gene transcription regulation


Accepted February 28, 2000